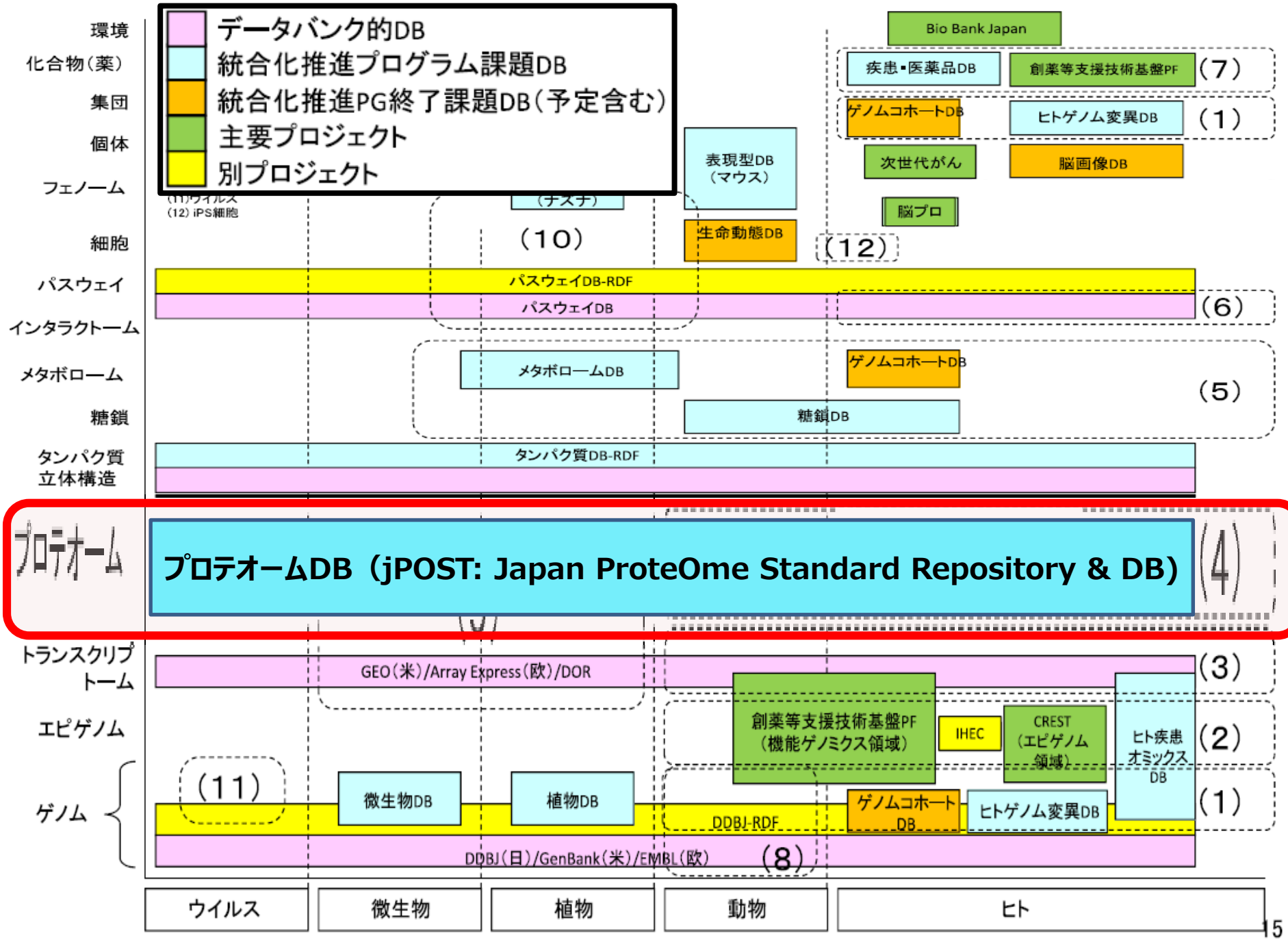
プロテオーム統合データベースの構築

代　表：石濱　泰（京大院薬）

主な共同研究者：五斗 進（京大化研）
荒木 令江（熊本大学）
松本 雅記（九州大学）
奥田 修二郎（新潟大学）
河野 信（DBCLS）


データバンク的DB
統合化推進プログラム課題DB
統合化推進PG終了課題DB（予定含む）
主要プロジェクト
別プロジェクト
環境
化合物（薬）
集団
個体
フェノーム
細胞
パスウェイ
インタラクトーム
メタボローム
糖鎖
タンパク質
立体構造
プロテオーム
トランスクリプトーム
エピゲノム
ゲノム
Bio Bank Japan
疾患・医薬品DB
創薬等支援技術基盤PF
(7)
ゲノムコホートDB
ヒトゲノム変異DB
(1)
次世代がん
脳画像DB
脳プロ
表現型DB（マウス）
生命動態DB
(10)
(12)
パスウェイDB-RDF
パスウェイDB
(6)
メタボローム DB
ゲノムコホートDB
(5)
糖鎖DB
タンパク質DB-RDF
プロテオームDB（jPOST: Japan ProteOme Standard Repository & DB)
(4)
GEO（米）/Array Express（欧）/DOR
(3)
創薬等支援技術基盤PF（機能ゲノミクス領域）
IHEC
CREST（エピゲノム領域）
ヒト疾患オミックスDB
(2)
(11)
微生物DB
植物DB
DDBJ-RDF
ゲノムコホートDB
ヒトゲノム変異DB
(1)
DDBJ（日）/GenBank（米）/EMBL（欧）
(8)
ウイルス
微生物
植物
動物
ヒト

プロテオーム解析からのみ得られる情報

- 翻訳後修飾
- 発現プロファイル（臓器、組織、細胞内）
- 絶対発現量
- タンパクータンパク相互作用

タンパク質は、発現調節（転写後調節）、局在、翻訳後修飾、分解、タンパク相互作用等によって、ダイナミックに姿を変え、場所を変え、消えたり現れたりを繰り返す。

タンパク質は生物機能発現の立役者！
生物現象や疾病のマーカーや薬の標的となる。

プロテオームはLS統合DBに欠けてはいけない必須情報

研究開発の背景

プロテオームDB: 世界の情勢

- HPP-HUPO ⟹ 遅々として進まず
- ProteomicsDB（LCMSデータ） ⟹ Nature, 2014
- Human Protein Atlas（ヒト抗体ライブラリー） ⟹ Science, 2015

Journal of proteome research
dx.doi.org/10.1021/pr500572z | J. Proteome Res. 2014, 13, 3854−3855
Letter
pubs.acs.org/jpr

Analyzing the First Drafts of the Human Proteome

Iakes Ezkurdia,[†] Jesús Vázquez,[§] Alfonso Valencia,[‡] and Michael Tress*[,‡]

> The results of our analysis show that both studies are substantially overestimating the number of protein coding and noncoding genes they find. We suggest that the experimental data from these two should be used with great caution, and we feel that these two unique draft maps of the human proteome should be put on hold until they can be carefully analyzed.

Nature 2014, the Human Proteome

Nature. 2014 , DOI: 10.1038/nature13302, PMID: 24870542

A draft map of the human proteome

Min-Sik Kim; Sneha M Pinto; Derese Getnet; Raja Nirujogi; Srikanth S Manda; Raghothama Chaerkady; Anil K Madugundu; Dhanashree S Kelkar; Ruth Isserlin; Shobhit Jain; Joji K Thomas; Babylakshmi Muthusamy; Pamela Leal-Rojas; Praveen Kumar; Nandini A Sahasrabuddhe; Lavanya Balakrishnan; Jayshree Advani; Bijesh George; Santosh Renuse; Lakshmi N Selvan; Arun H Patil; Vishalakshi Nanjappa; Aneesha Radhakrishnan; Samarjeet Prasad; Tejaswini Subbannayya; Rajesh Raju; Manish Kumar; Sreelakshmi K Sreenivasamurthy; Arivusudar Marimuthu; Gajanan J Sathe; Sandip Chavan; Keshava K Datta; Yashwanth Subbannayya; Apeksha Sahu; Soujanya D Yelamanchi; Savita Jayaram; Pavithra Rajagopalan; Jyoti Sharma; Krishna R Murthy; Nazia Syed; Renu Goel; Aafaque A Khan; Sartaj Ahmad; Gourav Dey; Keshav Mudgal; Aditi Chatterjee; Tai-Chung Huang; Jun Zhong; Xinyan Wu; Patrick G Shaw; ... (22 more)

The availability of human genome sequence has transformed biomedical research over the past decade. However, an equivalent map for the human proteome with direct measurements of proteins and peptides does not exist yet. Here we present a draft map of the human proteome using high-resolution Fourier-transform mass spectrometry. In-depth proteomic profiling of 30 histologically normal human samples, including 17 adult tissues, 7 fetal tissues and 6 purified primary haematopoietic cells, resulted in identification of proteins encoded by 17,294 genes accounting for approximately 84% of the total annotated protein-coding genes in humans. A unique and comprehensive strategy for proteogenomic analysis enabled us to discover a number of novel protein-coding regions, which includes translated pseudogenes, non-coding RNAs and upstream open reading frames. This large human proteome catalogue (available as an interactive web-based resource at http://www.humanproteomemap.org) will complement available human genome and transcriptome data to accelerate biomedical research in health and disease.

17,294 gene products

Nature. 2014 , DOI: 10.1038/nature13319

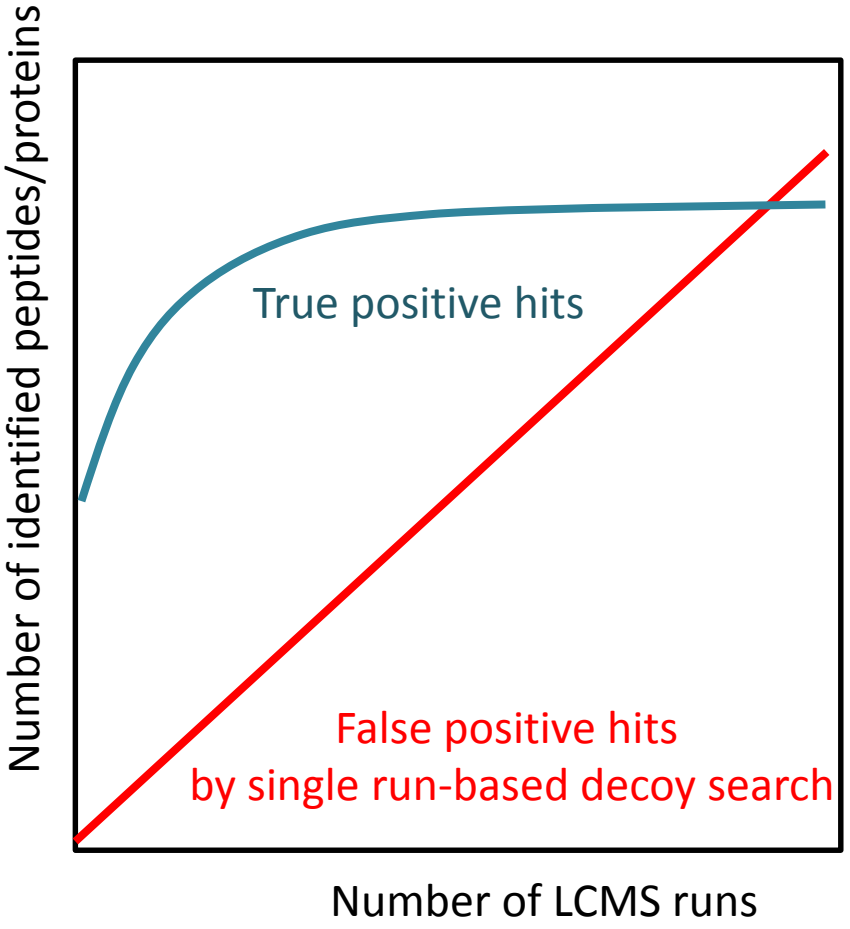
Mass-spectrometry-based draft of the human proteome

Mathias Wilhelm; Judith Schlegl; Hannes Hahne; Amin Moghaddas Gholami; Marcus Lieberenz; Mikhail M. Savitski; Emanuel Ziegler; Lars Butzmann; Siegfried Gessulat; Harald Marx; Toby Mathieson; Simone Lemeer; Karsten Schnatbaum; Ulf Reimer; Holger Wenschuh; Martin Mollenhauer; Julia Slotta-Huspenina; Joos-Hendrik Boese; Marcus Bantscheff; Anja Gerstmair; Franz Faerber; Bernhard Kuster

Proteomes are characterized by large protein-abundance differences, cell-type- and time-dependent expression patterns and post-translational modifications, all of which carry biological information that is not accessible by genomics or transcriptomics. Here we present a mass-spectrometry-based draft of the human proteome and a public, high-performance, in-memory database for real-time analysis of terabytes of big data, called ProteomicsDB. The information assembled from human tissues, cell lines and body fluids enabled estimation of the size of the protein-coding genome, and identified organ-specific proteins and a large number of translated lincRNAs (long intergenic non-coding RNAs). Analysis of messenger RNA and protein-expression profiles of human tissues revealed conserved control of protein abundance, and integration of drug-sensitivity data enabled the identification of proteins predicting resistance or sensitivity. The proteome profiles also hold considerable promise for analysing the composition and stoichiometry of protein complexes. ProteomicsDB thus enables navigation of proteomes, provides biological insight and fosters the development of proteomic technology.

18,097 gene products

ProteomicsDBの問題点とjPOSTの戦略



冗長性の高いMSデータをやみくもにかき集めた結果、多くの偽ヒットがProteomicsDBに登録。

MS解析データの標準化が必須

- 世界標準リポジトリ（PX)の採用
- 研究機関・プロジェクト毎に異なるフォーマットや解析法、信頼度評価法やアノテーションの標準化。
- 全データに対する統一したフィルタリング

多彩な生物種・翻訳後修飾・絶対発現量も含めた世界初の横断的統合プロテオームDB

jPOST（Japan ProteOme STandardization Repository & Database)

ProteomicsDB; self-corrected

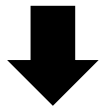
Molecular & cellular proteomics : MCP. 2015 , DOI: 10.1074/mcp.M114.046995, PMID: 25987413

A scalable approach for protein false discovery rate estimation in large proteomic data sets.

Mikhail M Savitski; Mathias Wilhelm; Hannes Hahne; Bernhard Kuster; Marcus Bantscheff

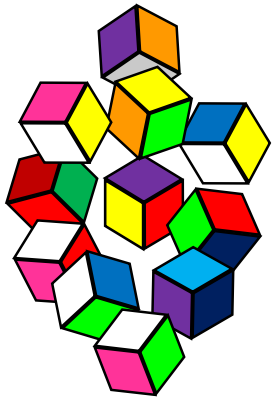
Calculating the number of confidently identified proteins and estimating false discovery rate (FDR) is a challenge when analyzing very large proteomic datasets such as entire human proteomes. Biological and technical heterogeneity in proteomic experiments further add to the challenge and there are strong differences in opinion regarding the conceptual validity of a protein FDR and no consensus regarding the methodology for protein FDR determination. There are also limitations inherent to the widely used classic target-decoy strategy (TDS) that particularly show when analyzing very large data sets and that lead to a strong over-representation of decoy identifications. In this study, we investigated the merits of the decoy-based protein FDR estimation approach taking advant collection comprised of ≈19,000 LC-MS/MS runs deposited i www.proteomicsdb.org). The "picked" protein FDR approach the same protein as a pair rather than as individual entities a decoy sequence depending on which receives the highest sc of this approach in combination with q-value based peptide instrument and search engine-specific differences. The "pick best when protein scoring was based on the best peptide q- stable number of true positive protein identifications over a demonstrate that this simple and unbiased strategy eliminat commonly used, "classic" protein FDR approach that causes protein identification in large data sets. The approach scales without losing performance, consistently increases the number of true positive protein identifications and is readily implemented in proteomics analysis software.

18,097 proteins (original)

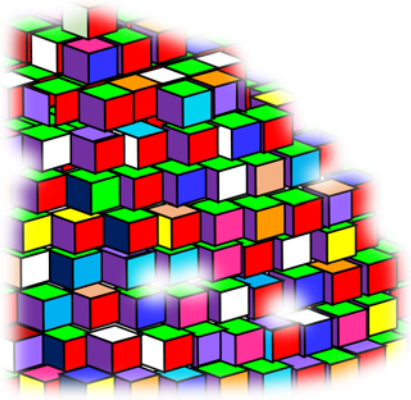


for this difference (supplemental Figure 7). We next applied the described data analysis strategy to the subset of data stored in proteomicsDB corresponding to our earlier publication on a mass spectrometry based draft of the human proteome (9). Using the classic FDR strategy 14,035 proteins were observed at 1% protein FDR compared to 14,714 proteins using the picked strategy. Applying the picked strategy without any protein score threshold yielded 17,326 proteins of the target database at 11.3% protein FDR corresponding to 15,290 true positive protein identifications in the dataset. When analyzing the complete current content of proteomicsDB (including the data of the Pandey proteome (10) and a number of further datasets), the number of protein identifications at 1% FDR increased to 14,638 (classic) and 15,375 proteins (picked) respectively.

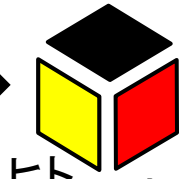
jPOSTの構成（ルービックキューブ型DB）



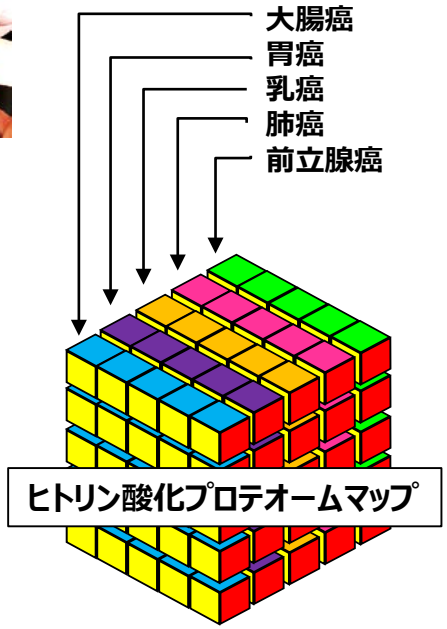
各プロジェクト毎のDB
「Cube」



Cubeを集積した全DB
「Globe」



間引きとアライメント



Globeから抽出・調整した
「Slice」

フォーカスドDB
（Slice）

集積DB
（Globe）

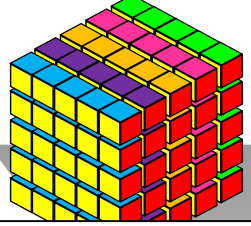
プロジェクトDB
（Cube）

標準化

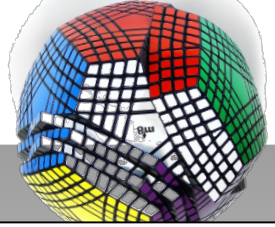
リポジトリ

測定生データ

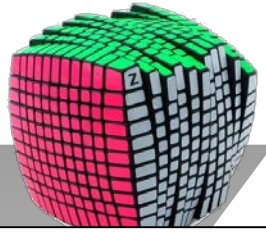
試料



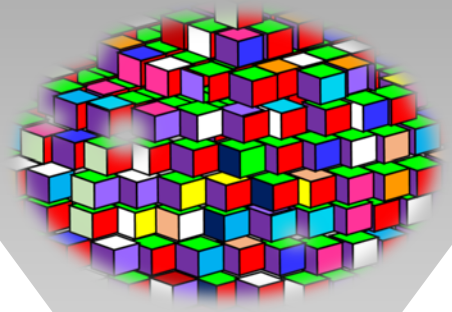
疾患別ヒトリン酸化
プロテオームマップ



生物種別代謝酵素PTMs
絶対発現量マップ



ユーザーカスタムマップ



例えば：ヒトとマウスの疾病別の転写因子群の発現量、リン酸化、アセチル化を見たい

抗体

質量分析

電気泳動

臨床検体

植物

モデル生物

培養細胞

細菌

希少生物

jPOSTの特徴

- データ標準化ステップを含む。
- ProteomeXchange コンソーシアムに参加。

  (開発費の抑制、データソースの確保、持続性)
- 段階ごとのDBシステム
- カスタマイズ可能なSlice DB
- RDFベース（将来の統合へ向けて）
- 翻訳後修飾および絶対量情報を含むプロテオームDB
- ヒトだけではなく、多彩な生物種も網羅。
- 国内・国際連携（JHUPO, AOHUPO, HUPO）

研究開発体制

代　表：石濱　泰（京大院薬）

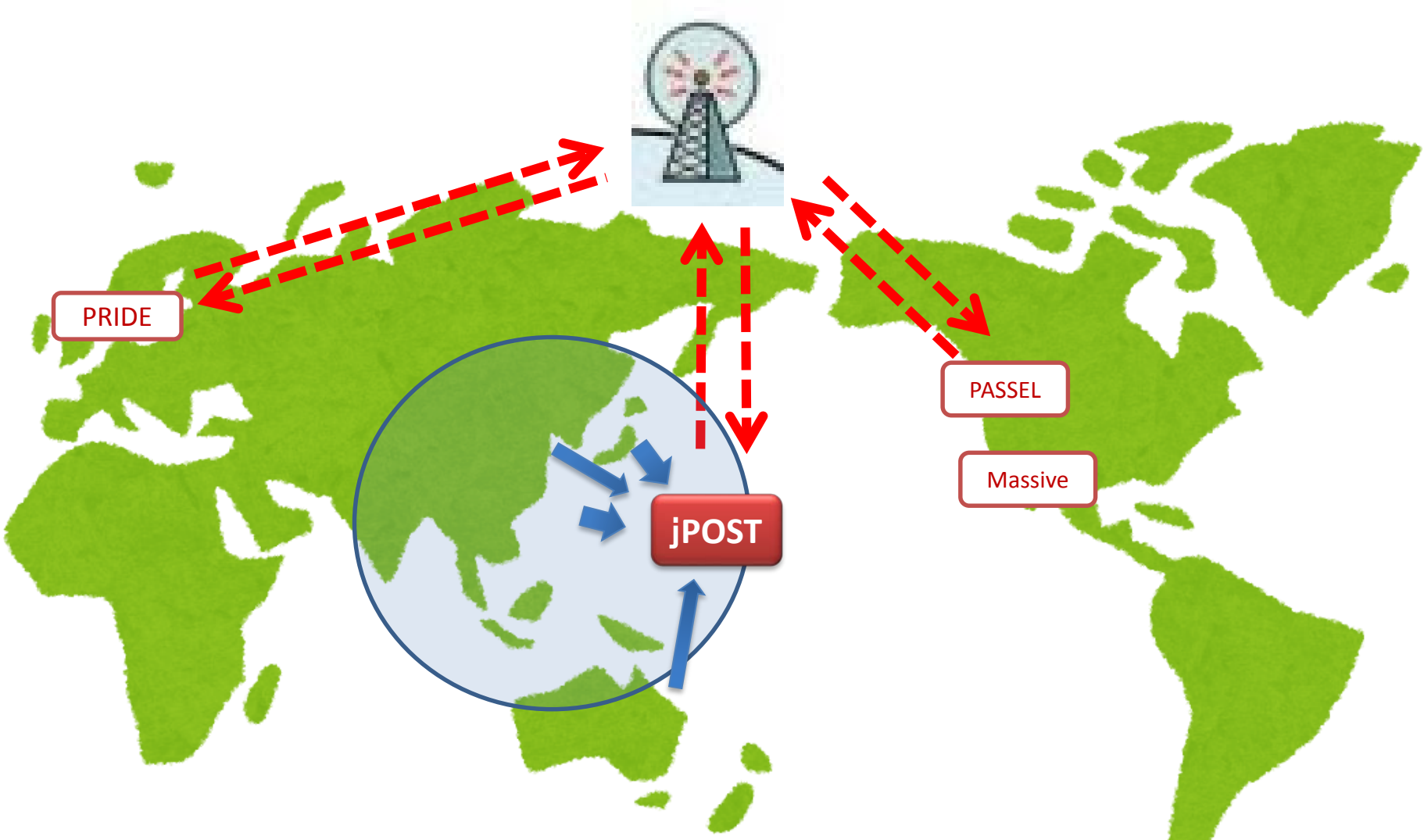
主な共同研究者：

- 五斗 進（京大化研）
- 奥田 修二郎（新潟大学）
- 河野 信（DBCLS）
- 荒木 令江（熊本大学）
- 松本 雅記（九州大学）

連携研究者　＆　保有データ：

連携研究者	所属	アドバイザー	所属
朝長　毅	医薬基盤研、JHUPO 会長、HUPO・AOHUPO 理事	平野　久	横浜市立大学
		戸田　年総	横浜市立大学
中神　弘史	理化学研究所 環境資源科学研究センター	村井　純	慶應義塾大学
近藤　格	国立がん研究センター	成松　久	産業総合技術研究所, 統合化プログラム糖鎖 DB (JCGGDB) 研究代表者
尾野　雅哉	国立がん研究センター		
中林　潤	横浜市立大学		

ProteomeXchange Consortium



アジアオセアニア地区唯一のリポジトリサイト

研究開発計画

項目	担当者	H27	H28	H29
サーバー管理・運用	**五斗**、河野	開発サーバー立ち上げ	定期的なデータ移行（開発→運用サーバー、定期的なシステム増設）	
		運用サーバー導入		
リポジトリ	**河野**、荒木、奥田	レポジトリシステム（PXC)導入	コアデータを用いたカスタマイズ	大規模データを用いた検証
		ストレージ導入	データ投稿システムの開発	外部公開に向けた検討	外部公開
標準化	**石濱**、松本、荒木、河野	ワークフローの設計（プラットフォーム別）	プロセスシステム開発	プロセス自動化
				統合化対応
プロジェクト別DB（Cube）構築	**松本**、石濱、荒木、五斗	Cubeの設計	グローバルプロテオミクスDB開発	ターゲットプロテオミクス開発
			PTMプロテオミクスDB開発	抗体・電気泳動DBの開発
集積キューブ型DB(Globe)構築	**奥田**、荒木、石濱、松本	RDFスキーマ、オントロジー設計	Globeの設計	Globeの作製
				統合化対応
フォーカスドDB（SliceMap）構築	**荒木**、河野、奥田、石濱、松本、五斗		スライスシステムの開発	プリセットスライスマップの作製
			ビューワーの開発	カスタムマップ用ツール開発

4月以降の活動

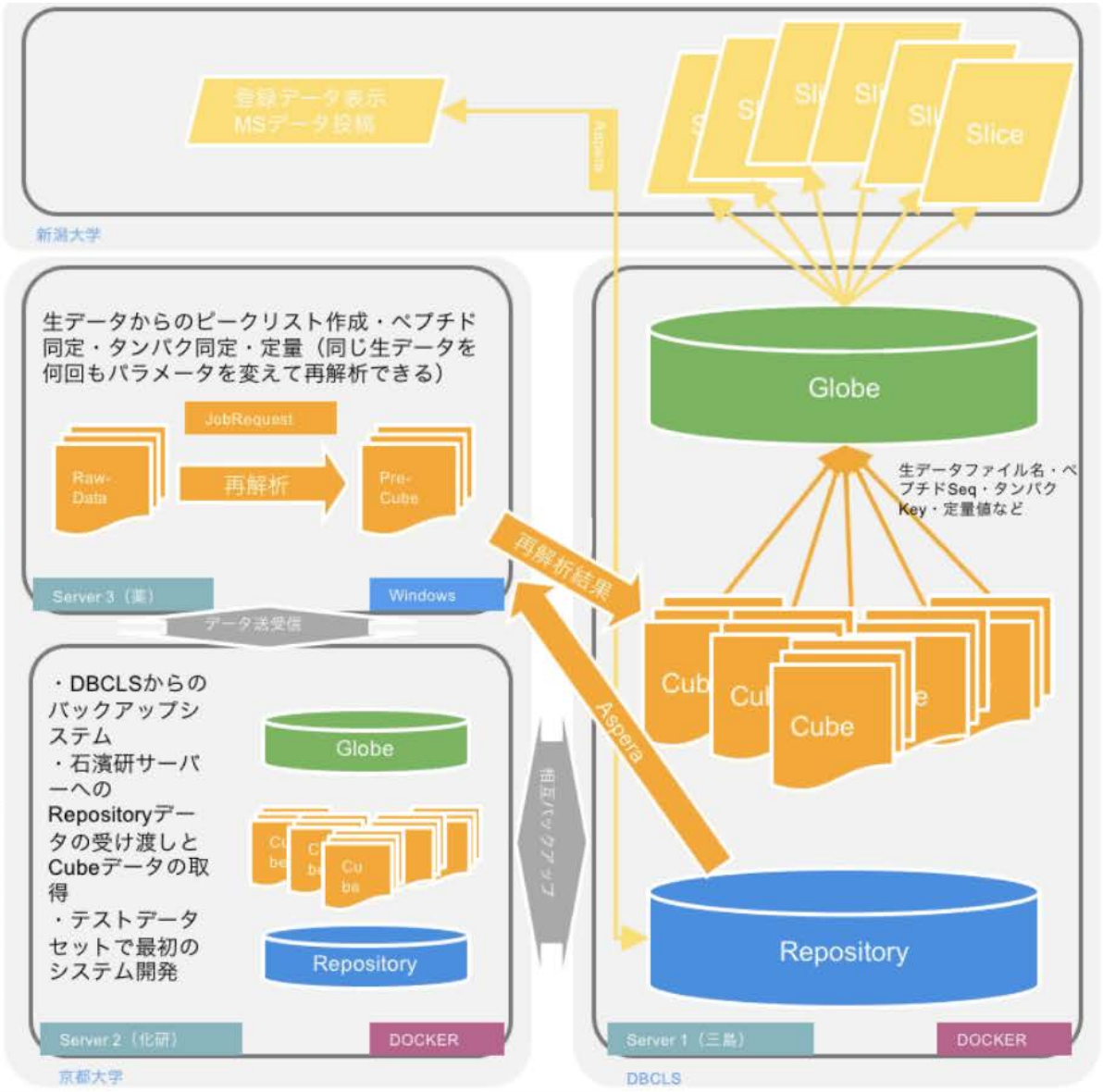
- 全体ミーティング
  - 4/7 (キックオフ)
  - 4/22 (WebEx)
  - 5/8 (WebEx)
- インフォチームミーティング（随時）
- サイト相互訪問
- EBI訪問（PXC-PRIDE）
- 拡大ミーティング（7月22日―熊本）
- HUPO2015（バンクーバー）
- 統合の日
- ワークショップ（10月13日―JST本部）

jPOST DBオーバービュー改訂版

ProteomeXchange
ProteomeCentral
EBI

連携PX ID
ットルデータAPI

（参考）申請書上の役割分担とスケジュール



ユーザー

ウェブインターフェイス
データ可視化
Mass++拡張

データ統合

SQLデータ管理
RDF化・SPARQL
オントロジー対応

キュレーション

ピークピッキング
ペプチド同定
タンパク同定・定量

データ登録

MS登録データベース/ツール
PX IDの取得

将来展望

- 本プロジェクトでは、UniProtのような寄せ集めタイプのプロテオーム知識ベースではなく、実験データをただやみくもに集めたProteomicsDBの失敗の教訓を活かし、国際的にもユニークな日本発の高質・高機能かつ多視点の統合プロテオームデータベースjPOSTの構築を目指す。
- 将来的には、本DBとDBCLSとの連繋を強化して、連邦型統合ライフサイエンスDBに必須のプロテオーム情報の供給源として貢献する。
- プロテオミクスを専門とする情報科学者を本プロジェクトを介して育成し、今後永続的なプロテオームDB維持のみならず、これらの情報を介して、生命科学、医薬分野における基礎および応用学の進歩に大きく貢献できる多くの人材の輩出に貢献する。