

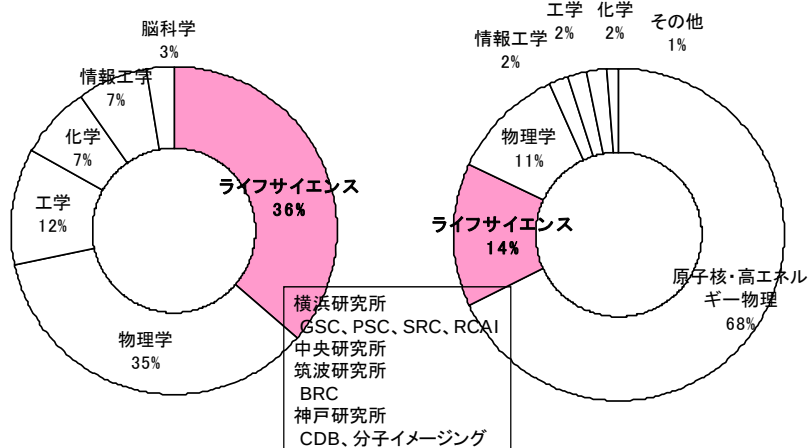


バイオ研究者向け支援サービス

独立行政法人理化学研究所
情報基盤センター 横浜チーム



スパコン利用者から見る バイオインフォマティクス概要



RSCC (スパコン) 利用のユーザ比率

研究分野ごとの利用時間の割合
(スパコン PC クラスタ部分)



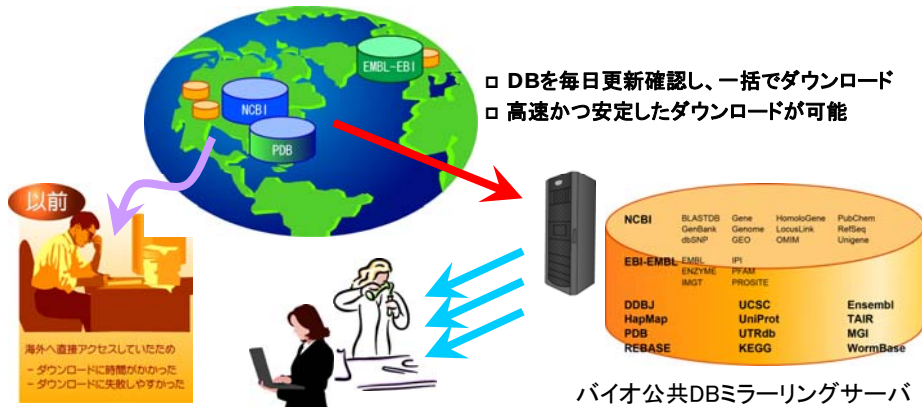
バイオ向けインフラ概要

- バイオ研究者向け支援サービス
- バイオ公共DBミラーリングサーバ(公共DB)
 - バイオ系基盤インフラの要
- RIKEN Super Combined Cluster(RSCC)
 - 大量ジョブ投入ユーザ向き大型計算リソース
- Bio Portal
 - 公共DB、RSCC、横浜SMPのサービスをラップしたWebアプリケーション
 - どちらかというとwet研究者・技術者向け
 - GUIを採用したBioinformatics初心者向き



バイオ公共DBミラーリングサーバ (公共DB)

- NCBI,EMBL-EBI,PDB等で公開されているバイオ系データベースの理研内部向けミラーリングサーバ
- 2003年8月サービス開始
- [ftp://yicdb.yokohama.riken.jp/pub/](http://yicdb.yokohama.riken.jp/pub/)





公共DB – 公開DB一覧

2008年1月7日現在

□ バイオ系データベース

- 配列データベース
- 遺伝子発現データベース
- 代謝経路データベース
- 化合物データベース
- 相互作用データベース
- 文献情報データベース

無償で公開されている代表的なものはほぼ網羅

ftpダウンロード状況

2007年度(2007/04/01 - 2007/02/29)

合計ダウンロードファイル数 : 4,024,029 files,

合計ダウンロードファイルサイズ :

7,750,345,527,186 Bytes

(のべ数にて)

□ ミラーリングされている代表的なDB

DB名	データ容量	DB名	データ容量
BLASTDB(NCBI)	約118GB	genomes(NCBI)	約225GB
DDBJ Release	約139GB	GEO	約2376GB
DDBJ New	約14GB	GO	約2GB
EMBL Release	約100GB	goldenPath	約932GB
EMBL New	約8GB	KEGG	約159GB
Ensembl	約1029GB	PDB	約133GB
GenBank Release	約153GB	PFAM	約93GB
GenBank New	約3GB		



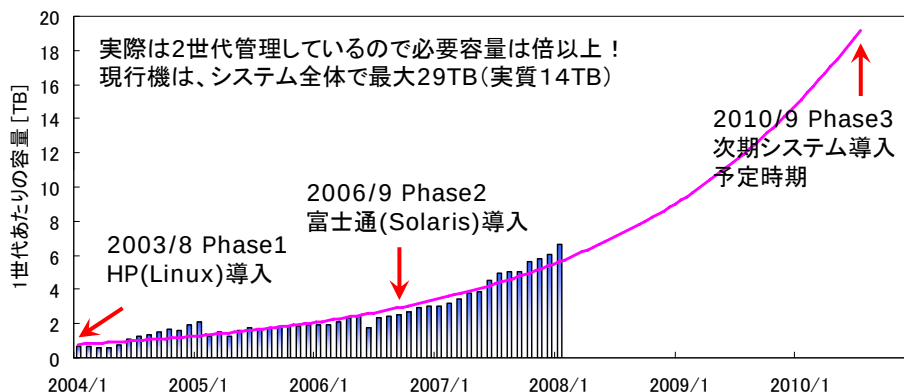
公共DB – ストレージ増設状況

□ ミラーリング状況

□ <http://yicdb.yokohama.riken.jp/public/status/>

□ DB容量の指数的な増加 ⇒ 空き容量の圧迫

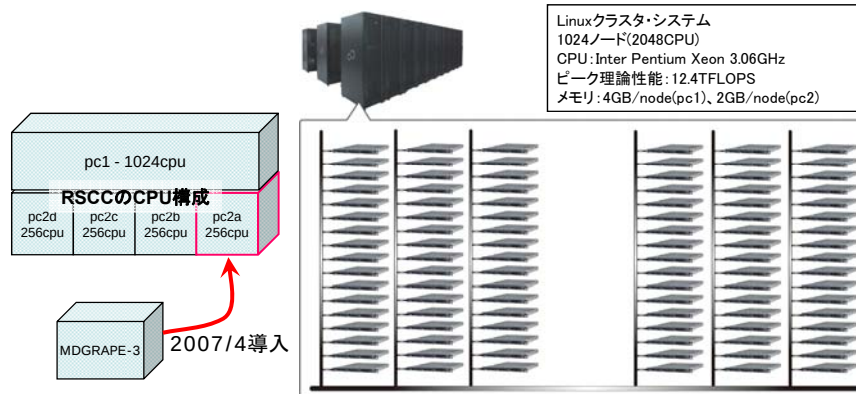
□ ディスクの追加を必要に応じて実施





RIKEN Super Combined Cluster (RSCC)

- Linuxクラスタとベクトル計算機SX-7から構成される共用計算機
- 2004年4月から現行機サービス開始
- <http://ccr.riken.jp/rsccl/index.html>
- Linuxクラスタ部は1024ノードから構成
- 1CPUジョブ、バイオ系アプリ(分散型ジョブ)はpc2a



RSCCにインストールされているアプリケーション

- BLAST 2.2.10
 - /usr/local/BIO/blast
- Hi-per BLAST
 - /opt/hiperblast/bin/hiperblast
- FASTA
 - /usr/local/BIO/FASTA
- HMMER 2.3.2
 - /usr/local/BIO/hmmer
- CLUSTALW 1.83
 - /usr/local/BIO/clustalw
- InterProScan 4.3.1
 - /usr/local/BIO/iprscan
- GOLD / Rmpi + Bioconductor / wise / RepeatMasker / MaskerAid
- MDGRAPE-3ライブラリ



RSCCですぐに利用できるDB

2008/1/31現在

□ /work/biodb/current/pub/

DB名	概要	公開元サイト	データ容量
BLASTDB	相同性検索用データベース	National Center for Biotechnology Information ftp://ftp.ncbi.nlm.nih.gov/blast/	約120GB
Ensembl	生物別ゲノム・タンパク質データベース (homo_sapiensなど一部種に限定)	The Ensembl project ftp://ftp.ensembl.org/pub/	約46GB
genomes(NCBI)	生物別ゲノムデータベース	National Center for Biotechnology Information ftp://ftp.ncbi.nlm.nih.gov/genomes/	約218GB
Pfam	タンパク質ドメインデータベース	The Sanger Institute ftp://ftp.sanger.ac.uk/pub/databases/Pfam/	約89GB
PDB	タンパク質構造データベース	The RCSB Protein Data Bank ftp://ftp.wwpdb.org/	約70GB
PROSITE	タンパク質配列モチーフデータベース	Expert Protein Analysis System ftp://ftp.expasy.ch/databases/	約4.2GB
Uniprot	アミノ酸配列データベース	The universal protein resource ftp://ftp.uniprot.org/pub/databases/	約56GB



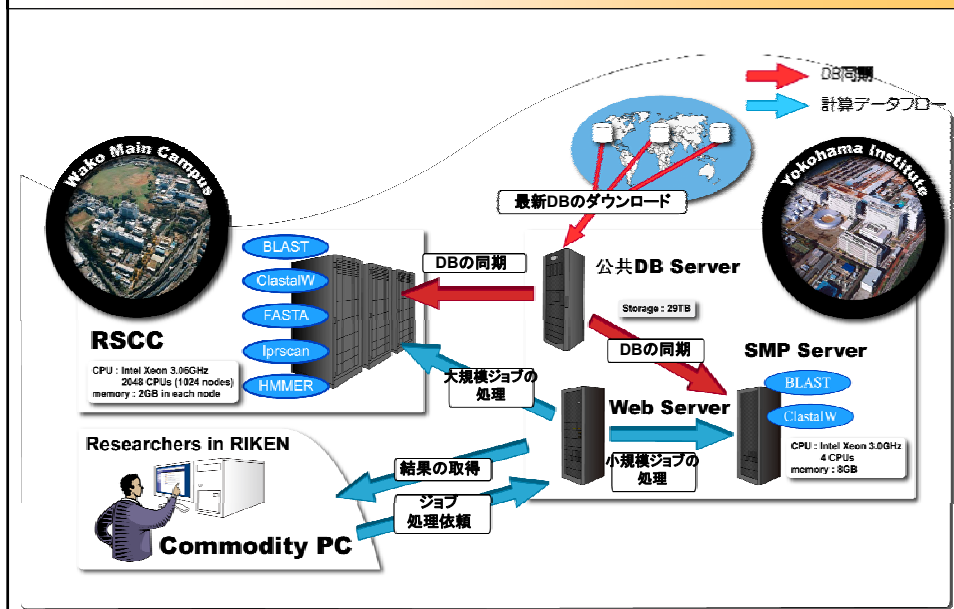
Bio Portal

- Bioinformatics入門者向けサービス
- 2005年3月からサービス開始
- ブラウザベースのGUI(Webアプリケーション)
- 公共DB、RSCCを意識せずに包括的な利用が可能
- アプリケーション
 - BLAST
 - FASTA
 - CLUSTALW
 - HMMER
 - InterProScan



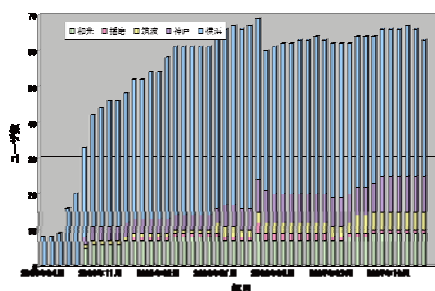


Bio Portal システム概要



Bio Portal システムの特徴

- 全理研から利用可能
(2008/3現在 ユーザ登録数63名)
- 利用するDBサイズによって
計算リソース(RSCC/横浜
SMPサーバ)を振り分ける
アルゴリズム



H.Yuasa, T.Iwasaki, M.Kurokawa, T.Shigetani, T.Horiki, R.Himeno, *Development of User-friendly Supercomputing Portal in Bio Research Field, e-Science*, (2005).

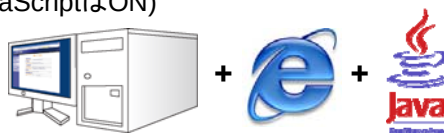
- RSCC / 横浜SMPサーバで利用するDBは公共DBから最新
のものをダウンロード
- browser-base のアプリケーションとして振舞う
- 高い処理能力とインフラ構築に必要なコストの節約が可能



Bio Portal クライアント動作条件

□ 動作条件

- 理研のネットワークに接続できるPC等
- 標準的なWebブラウザ (JavaScriptはON)
 - IE / Firefox / Safari
- JAVA 1.4.2_05以上



□ 利用申請

- <http://yicdb.yokohama.riken.jp/info/j-account.html>

□ ポータル画面トップ

- <https://yic-bio-web.yokohama.riken.jp/>
- 利用説明は、**Bio Portal「ご利用開始」ガイド**をご覧ください。



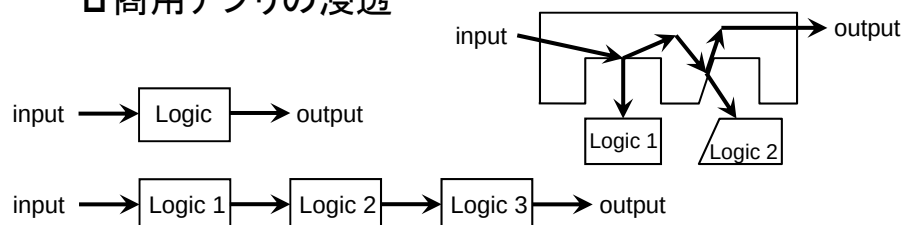
Bio Portal次期計画





状況確認

- 次期スパコンリプレースが2009年4月
- Bio Portalは来年度以降の開発計画
- 3年間でユーザの利用形態は多様化
 - データマイニング・可視化の多様化
 - ツールのパイプライン・ワークフロー化
 - 商用アプリの浸透



現行システムの機能一覧

- 機能単位
 - Interface
 - DB更新処理
 - ジョブステータス管理
 - 認証・ユーザ管理
 - アプリ・ロジック切り替え
 - リソース選択
 - 障害管理
- 密結合であることの利点と問題点
 - トランザクション
 - 開発・テスト工程数



Interfaceの問題点

- Interfaceと機能の実装が密結合
- ソフトウェア・パーツの寿命
 - ポータルの基幹部分とInterfaceのライフサイクルを同じにするとデータマイニングの陳腐化
 - サービスの停滞

- バイオ系研究者のアクティビティ

- Bioinformatics Link Directory

http://bioinformatics.ca/links_directory/

リンク数 1200 (2007年のUpdate数 130)

J.A.Fox, S.McMillan, B.F.F.Ouellette, Conducting Research on the Web: 2007 Update for Bioinformatics Links Directory, Nucleic Acid Research, Vol.35, No.Web Server issue W3-W5 (2007).



他の問題点

- システムの巨大化・複雑化
 - 保守コストの増大
 - ソフトウェア・パーツの再利用性の低下
 - ハードエラーよりもソフト(運用)エラーの方がはるかに発生しやすい...
- パッケージング
 - 組み立てキット? 完成品? ユーザの好みは...
 - 理研におけるユーザ層の広がり



現行システムの見直し

- 情報基盤センターはリソース提供者
- プログラムによるハンドリング
- ハードコードされたバインディングではなく、必要に応じて変更できるシステム
- サービスの粒度とユーザビリティ

- 複数のアプリケーション(サービス)の連携
- Interfaceの柔軟な変更



Webサービス化

- Webサービス
 - 主にHTTP上で動作しRPCやXMLでメッセージ交換を実現する技術
 - システム連携をする上で相互運用性が優れている技術
 - アウトソーシングの基本形
 - プログラム・スクリプト中からリソースを利用するための技術
- SOAP(RPC) と REST
 - RESTの利点
 - 軽量(Lightweight) – XMLが単純
 - 可読性が高い
 - 開発が容易 – GET / POST / PUT / DELETE
 - SOAPの利点
 - (プログラマ的に) 使いやすいこと
 - W3Cによる規格(プロトコル)であること
 - 開発ツールが多いこと(マルチベンダー化)



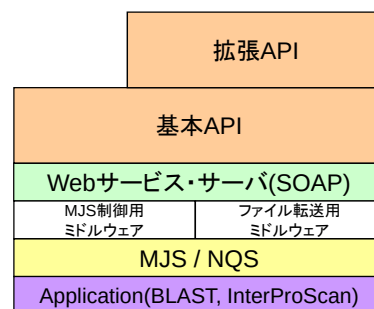
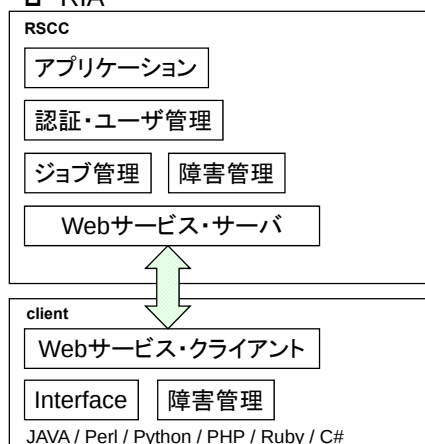
バイオ系サイトのWebサービス

- DDBJ(DNA Databank of Japan)
 - http://www.xml.nig.ac.jp/index_jp.html
 - Web API for Bioinformatics (WABI)
- EBI(European Bioinformatics Institute)
 - <http://www.ebi.ac.uk/Tools/webservices/>
- NCBI(National Center for Biotechnology Information)
 - http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html
 - Entrez Programming Utilities
- KEGG(Kyoto Encyclopedia of Genes and Genomes)
 - <http://www.genome.jp/kegg/soap/>
- PathPort ToolBus tools / BioMOBY / BIND



次期システム

- Webクライアントの1つとしてBio Portal を運用
- アプリケーションのワークフロー・パイプラインはクライアント増設で対応
- 複数年度で計画
- RIA



第1段階: シンプルなトランザクション
第2段階: Bio PortalのWebクライアント化
第3段階: 特定のグループとの相互運用



課題

- スループットの保持
- ファイル転送方法、大ペイロード対策
- ユーザデータの扱い
- 横浜側リソースとの連携
- 機能ごとに開発されたソフトウェア・パーツの再利用性の維持
- マルチベンダー開発時の責任分解点



まとめ

- 公共DB
 - ストレージ容量強化
 - テープ装置とディスク装置の透過的な運用
 - テープ装置の見直し
- RSCC
 - 2009年4月にリプレイス予定
- Bio Portal
 - Webサービス化
 - データマイニング・可視化方法の見直し