

統合データベース支援： バイオDBサーバー構築演習

森下 真一
中谷 洋一郎

目的

- バイオDBを構築できる人材を育てる
 - 膨大なソフト外注費(150～200万円/月)を回避
 - DBの保守・拡張が自前でできること
 - やむをえず外注する場合も、正確な仕様書を書ける力と、納入されたソフトの問題点を見抜く力を養う
- 必要スキルを1年間のカリキュラムで教え込む
- 次の1年で独創的サーバーを構築

計画

DB 構築者を養成するために以下の3つの演習を実施する。

① バイオ DB サーバー構築演習

データベースサーバーのミラーサイトを構築する。OS, apache, MySQL 等の主要ソフトウェアのインストールおよびネットワークセキュリティに習熟することが目標である。参加者には各自にサーバー構築用ワークステーションを配布する。演習を完了するまでには、受講者の能力と受講可能時間に応じて最短で3ヶ月、最長で1年間の時間を予定している。

② プログラミング演習

Java および Perl プログラミングを演習した後に、アルゴリズムの知識を活かした配列処理やデータマイニングの実装を行う。上記①バイオ DB サーバー構築演習では実施がむずかしいプログラミング演習を行うことで、独自にソフトウェア構築ができる能力を身につけることをめざす。演習総時間は90時間で約2ヶ月間を予定している。

③ 独創的サーバー構築演習

大規模計算のためのクラスター利用技術を習得させ、他に類の無いバイオDBサーバーを設計、実装、公開することを目標とする。バイオDBサーバー構築演習およびプログラミング演習を修了した受講者に対して平成20年度より開講を予定しており、そのための計算機セットアップを平成19年度に準備した。

年次計画

平成19年度

20年度

21年度

22年度

プログラミング演習
(夏季 90時間)

バイオDB
サーバー構築演習
(通年 毎週演習)

註)教育プログラムを早期に
立ち上げるため、2007年度
に限ってはプログラミング演習
とバイオDBサーバー演習を並
行実施

プログラミング
経験者

プログラミング演習
(夏季 90時間)

バイオDB
サーバー構築演習
通年 毎週演習 約9名
1ヶ月間 短期演習 約1名

独創的サーバー
構築演習
通年の課題 5名

第1期生(5名)

プログラミング
経験者

註)プログラミング演習が不
要と判定されたプログラミング
経験者はバイオDBサーバー
構築演習に進むことができる

バイオDB
サーバー構築演習
通年 毎週演習 約3名
1ヶ月間 短期演習 約2名

独創的サーバー
構築演習
通年の課題 10名

第2期生(10名)

独創的サーバー
構築演習
通年の課題 5名

第3期生(5名)

演習用WS15台
(平成19年度予算申請)

註) 1期生と2期生が20年度には重なること(21年度は2, 3期生)、WSが15台であること、
演習スタッフ1.5名による徒弟制度であるため、各年15名の受け入れが限度である

平成21年度受講者

- 一年コース(4月～3月)
 - 東大情報生命科学専攻から3名
 - 東大医学系研究科から1名

DBサーバー構築演習の目標設定

- 1: CentOS を自分のマシンにインストールする
- 2: ネットワークと接続する
- 3: セキュリティアップデートを行う
- 4: Web サーバーを立てる(ファイヤーウォールの設定を行う)
- 5: CGIを設置してみる
- 6: MySQL サーバーを立てる
- 7: 簡単なデータベースを作成する
- 8: Ensembl core をインストールしミラーを作成する
- 9: 複数種の実データをダウンロードして完全ミラーを作る
- 10: バックアップを作成して即時復旧できる体制を作る

21年度バイオDBサーバー構築演習の概要

	演習日程	テーマ
• OS (Linux) のインストール	• 4/16	イントロダクション、CentOSのインストール
• ネットワーク・ファイアーウォールの設定	• 4/23	セキュリティと定期アップデート、SSHによる外部からの安全な接続
• Webサーバーの設置・設定 (apache)	• 4/30	Webサーバーの設置、シェルスクリプト、Pukiwikiの設置
• RDBMSの設置・設定 (MySQL)	• 5/07	Perl演習
• Perl モジュールの設置・設定	• 5/14	CPANを使いこなす、BioPerlのインストール
• Ensembl の設置・設定	• 5/21	RDBMS、Perlからデータベースを扱う
• Perl, Javaプログラミング	• 5/28	PerlによるCGIプログラミング
• CGIからのデータベース検索	• 6/04	Java演習: プログラムの書き方
• メンテナンス全般	• 6/11	Java演習: データ構造とオブジェクト
– 障害対応	• 6/25	Java演習: GUIアプリケーションとデータの入出力
– ソフトウェアの Security fix やバージョンアップ等	• 7/02	Java演習: データベースアプリケーション
	• 7/09	CGIでデータベースを検索する
	• 7/16	Ensemblデータベースをミラーする1
	• 7/30	Ensemblデータベースをミラーする2
	• 9/03	Ensemblデータベースをミラーする3
	• 9/17	Ensemblデータベースをミラーする4
	• 10/01	Ensemblデータベースをミラーする5
	• 10/15	サーバーのバックアップ1
	• 10/15	BLATを用いたmRNAのゲノムへのマッピング
	• 10/29	サーバーのバックアップ2
	• 10/29	Ensemblデータの解析、BioMartを使ったデータ取得
	• 11/12	サーバーのバックアップ3
	• 11/26	サーバーのバックアップ4
	• 11/26	OpenCVを使った画像処理演習
	• 12/10	UTGB Toolkitのインストール
	• 12/10	JFreeChartを使用したグラフの描画
	• 12/10	遺伝子発現データベースを使い倒す
	• 12/24	UTGB Toolkitを使ったゲノムブラウザプログラミング
	• 12/24	遺伝子発現データの生物学的な解釈

OSのインストール

- 講義日程: 4/16
- システム・ネットワーク・ウェブ・データベース等に関する基礎的な用語の解説。
- 各自のサーバーにLinuxをインストール。
 - CDイメージをダウンロードしCentOS最新版をインストールする。

セキュリティと自動更新設定 鍵認証方式によるログイン

- 講義日程: 4/23
- セキュリティーについて。
 - 脆弱性とは？
 - 脆弱性の例。
 - Buffer overflow, SQL injection, Cross site scripting, Brute force attack, DNS spoofing, ...
- yum-cronによる定期的なセキュリティアップデートの設定。
- ネットワークの設定。
- SSHの設定。
 - 公開鍵認証方式によるログイン。
 - パスワードを入力しない安全な方式で外部からssh接続を行う。

Web サーバーの設置

Pukiwikiの設置

- 講義日程 : 4/30
- ウェブサーバーの設置。
 - Apacheのインストール。
 - 設定ファイルの編集。
 - Firewallの設定。
- Pukiwikiの設置。
 - ウェブ上で情報の共有と整理を多人数で行える。
 - Pukiwikiをダウンロードし、サーバーにインストールする。
 - Pukiwikiの基本操作、文法の解説。

Perlプログラミング演習

- 講義日程: 5/7
- なぜPerlを学ぶのか？
 - “バイオインフォマティクスの分野で、最も広く使われているスクリプト言語。”
 - Ensembl のコードもPerlで書かれているためミラーサイト構築時にPerlの知識が必要。
- Perlのインストール。
- 基本的なPerl文法の解説。
 - File I/O, 正規表現, サブルーチン, ソートなど。
- ゲノム配列データをダウンロードし、Perlを使用して簡単なデータ処理を行う。

ソフトウェア・モジュールのインストール

- 講義日程: 5/14
- 他の研究者によって開発されたソフトウェア・ライブラリー・モジュールを使用することで、解析プログラム・解析パイプラインをすばやく簡単に作成することができる。
- CPANの利用。
 - CPAN の使い方 ライブラリ
 - CPAN (Comprehensive Perl Archive Network) とは何か
 - 最初の configuration
 - モジュールのインストール
 - 依存モジュールが足りない場合は
 - 自分のHOME下へのインストール
 - インストール済みモジュールのチェック
- CPANからPerlモジュールをインストールする。
 - AppConfig, DBI, DBD::SQLite, File::HomeDir, YAML, Spreadsheet::ParseExcel, Spreadsheet::WriteExcel, Cwd, SVG, PostScript::Simple, HTML::Parser, XML::Parser, IO::Zlib, Term::ReadLine, Template, Digest::SHA::PurePerl, Bundle::BioPerl
- makeによるモジュールインストール。

データベースの設置

Perlを使ったデータベース検索

- 講義日程 : 5/21
- データベースの設置。
 - MySQLのインストール。
 - MySQLの基本的なコマンドの解説。
 - データベースの検索。
- PerlのDBIモジュールによるデータベースアクセス。
 - Perlプログラムからの遺伝子データの検索。
 - BioPerlを使用した遺伝子系統樹解析。

PerlによるCGIプログラミング

- 講義日程 : 5/28, 7/9
- Perlを使ってCGIプログラムを作成
 - HTTPの解説。
 - Perlでアクセスカウンターを作成。
 - GET方式とPOST方式によるユーザーからの入力の処理。
 - Cookieの解説。
- ウェブページからユーザー入力を受け取りデータベースを検索するCGIの作成
 - BioPerlを使って、TreeFamデータベースの系統樹データを検索。
 - CGI作成用のPerlモジュール HTML::Template, HTML::FillInformを使用。

Java プログラミング演習

- 講義日程: 6/4, 11, 25, 7/2
- 演習内容
 - プログラムの書き方。
 - Javaの仕組みと文法、Eclipse (Javaの開発環境) の使い方
 - データ構造とオブジェクト。
 - 配列、オブジェクト指向プログラミング、データ構造
 - GUIアプリケーションとデータの入出力
 - 文字列、オブジェクト・クラス、入出力、GUIアプリケーションの作成
 - データベースアプリケーション
 - リレーショナルデータベースとSQL, SQLite JDBCを使ってJavaからデータベースを扱う。

Ensemblミラーサイトの構築 サーバーのバックアップ

- 講義日程 : 7/16, 30, 9/3, 17, 10/1, 15, 29, 11/12, 26
- Ensembl ミラーサイト構築
 - 必要なモジュールのインストール
 - データのダウンロードとインストール
 - Ensemblウェブサイトの設定、起動
- TeraStationへのバックアップ
 - データベースのバックアップ
 - "rsync"コマンドによるバックアップ

UTGB toolkitによる ゲノムブラウザ開発

- 講義日程：12/10, 24
- UTGB toolkitを用いて新しいタイプのゲノムデータをトラックに表示する技術を習得することが目標。
- UTGBの紹介、UTGB toolkitのインストール。
- UTGB toolkitを用いてデータを表示する。

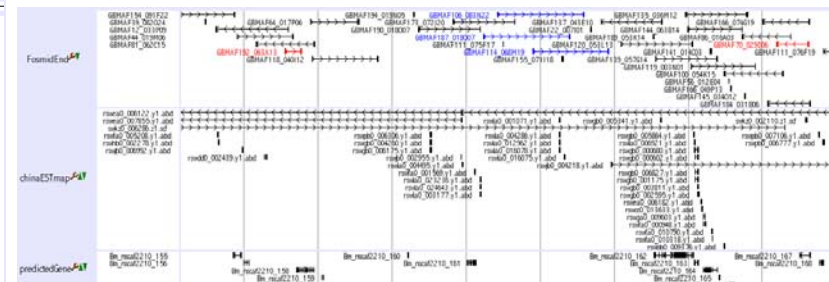
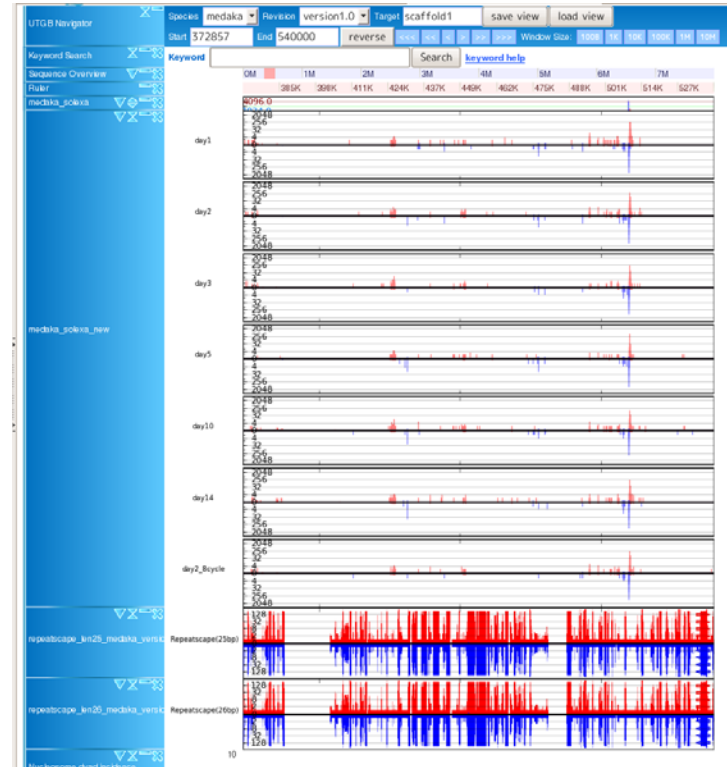
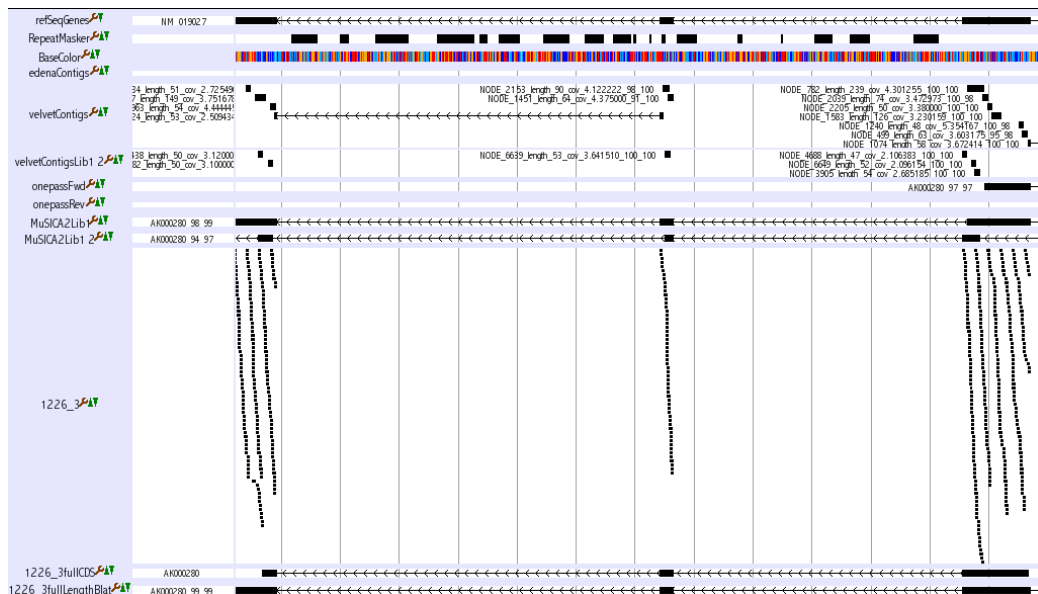
21年度講習の進捗

(受講者数:4名)

- ① 1. CentOS を自分のマシンにインストールする
- ② 2. ネットワークと接続する
- ③ 3. セキュリティアップデートを行う
- ④ 4. web サーバーを立てる(ファイヤーウォールの設定を行う)
- ⑤ 5. CGIを設置してみる
- ⑥ 6. MySQL サーバーを立てる
- ⑦ 7. 簡単なデータベース作成をする
- ⑧ 8. Ensembl core をインストールしミラーを作成する
9. 複数種の実データをダウンロードして完全ミラーを作る
- ⑨ 10. バックアップを作成して即時復旧できる体制を作る

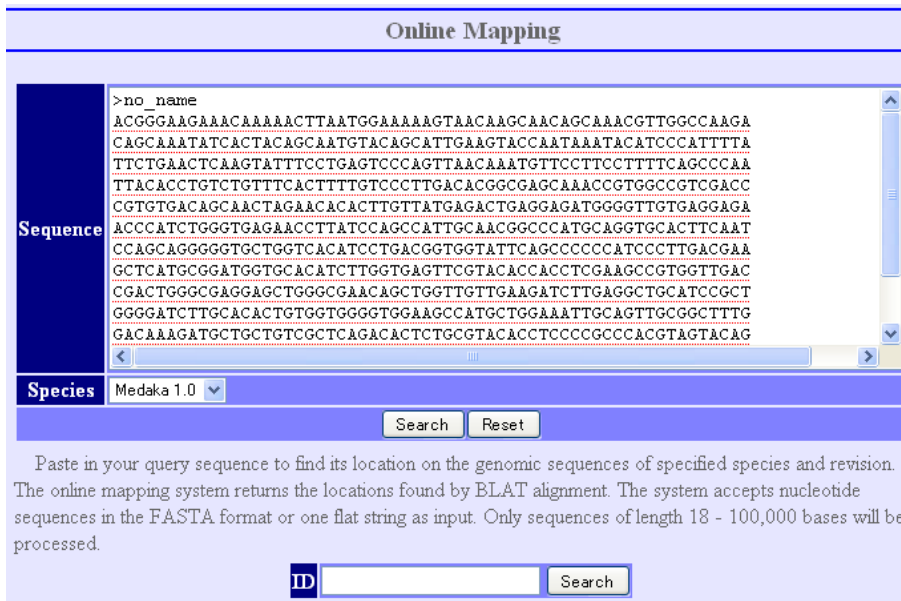
独創的サーバー構築演習

- 受講者が研究で使用する新規データをゲノムブラウザーに表示する。
 - 発現量データを表示するトラックの開発。
 - 配列特異性を視覚化するトラックの開発。
 - “RepeatScape”として公開。
 - Fosmid-end解析, 完全長cDNAアセンブリーの解析をブラウザーに表示。
- データ解析・論文作成に活用されている。



UTGB Medaka Online Mapping

- クラスターでアラインメントの計算
- ウェブブラウザでマッピング結果を表示



ID: 20090120175237_23233

Alignm	match	mis-match	rep. match	N's	Q gap count	Q gap bases	T gap count	T gap bases	strand	Q num
View	793	0	0	0	1	1	2	884	+	no_

>no_name:0+794 of 794 scaffold1211:611860+613537 of 1085344
ACGGGAAGAAAACAAAACTTAATGGAAAAAGTAACAAGCAACAGCAAAAGCTTGGCCAAAGA
|||||
ACGGGAAGAAAACAAAACTTAATGGAAAAAGTAACAAGCAACAGCAAAAGCTTGGCCAAAGA
CAGCAAAATATCACTACAGCAATGTACAGCATTGAAGTACCAATAAATACATCCCATTTTA
|||||
CAGCAAAATATCACTACAGCAATGTACAGCATTGAAGTACCAATAAATACATCCCATTTTA
TTCTGAACCTAAAGTATTTCTGAGTCCAGTTAACAATGTTCCTTTCCAGCCCAA
|||||
TTCTGAACCTAAAGTATTTCTGAGTCCAGTTAACAATGTTCCTTTCCAGCCCAA
TTACACTGTCTGTTTCACTTTTGCCCTTGACACGGCGAGCAAACCGTGGCCGTCGACC
|||||
TTACACTGTCTGTTTCACTTTTGCCCTTGACACGGCGAGCAAACCGTGGCCGTCGACC
CGTGTGACAGCAACTAGAACAACACTTGTATGAGACTGAGGAGATGGGGTTGTAGAGAGA
|||||
CGTGTGACAGCAACTAGAACAACACTTGTATGAGACTGAGGAGATGGGGTTGTAGAGAGA
ACCCATCTGGGTGAGAACCTTATCCAGCCATTGCAACGGGCCATGCAAGGTCACATCCGCT
|||||
ACCCATCTGGGTGAGAACCTTATCCAGCCATTGCAACGGGCCATGCAAGGTCACATCCGCT
GGGGATCTTGCACACTGTTGGTGGGGTGGAAAGCCATGCTGGAAATTGCAGTTGCGGCTTTG
|||||
GACAAAGATGCTGCTGTCGCTCAGACACTCTGCGTACACCTCCCGCCACAGTAGTACAG
-----80-
CCAGCAGGGGGTGTGGTGCATCTCAGCGGTGGTATTACAGCCCCCATCCCTAAGG...
-----CTTGACGAAGCTCATGGGATGGTGCACATCTTGGTGAAGTTCGTACACCACTC
|||||
TGTACTTGACGAAGCTCATGGGATGGTGCACATCTTGGTGAAGTTCGTACACCACTC
GAGCCCTGGTTGACCCACTGGCCAGGAGCTGGGGCAACAGCTGGTTTGAAGATCTT
|||||
GAGCCCTGGTTGACCCACTGGCCAGGAGCTGGGGCAACAGCTGGTTTGAAGATCTT
GAGGCTGCATCCGCTGGGATCTTGCACACTGTGGTGGGGTGGAAAGCCATCTGAAATT
|||||
GAGGCTGCATCCGCTGGGATCTTGCACACTGTGGTGGGGTGGAAAGCCATCTGAAATT
GCAGTTGCGGCTTTGGACAAGATGCTGCTGCGCTAGACACTCTGGGTACACTCCCC
|||||
GCAGTTGCGGCTTTGGACAAGATGCTGCTGCGCTAGACACTCTGGGTACACTCCCC
GCCACGTAGTACAGGTGTAACC-----804-----CTTGCCATATGCTGCGCGT
|||||
GCCACGTAGTACAGGTGTAACCCTGGGA...CTTACCTTTGCTATGCTGCGCGT
GTGCTCGATGGTGGAGTTGCGGTTGACCTTGGAAAGGAGGCCAGGCAAGAGCGTTCTT
|||||
GTGCTCGATGGTGGAGTTGCGGTTGACCTTGGAAAGGAGGCCAGGCAAGAGCGTTCTT
GTGTTGACAGGGGTCACTGAAGCCGCTCCACCAAAGATGCTGTGG
|||||
GTGTTGACAGGGGTCACTGAAGCCGCTCCACCAA--GATGCTGTGG

昨年度までのバイオDB構築演習

受講者による講義・演習

- BLATを用いたmRNAのゲノムへのアラインメントと出力結果の処理。
- Ensemblの比較ゲノムデータを使った解析とBioMartによるデータ取得方法。
- OpenCV(画像処理ライブラリー)を用いた画像解析。
- 遺伝子データベースを使い倒す(BioGPS, NCBI Gene Expression Omnibus, Mouse Genome Informatics データベースを使った発現解析)。
- 遺伝子発現データの生物学的な解釈(DAVID, Reactomeデータベースを使った発現解析)。

サーバー使用者氏名とネットワーク図

es1.gi.k.u-tokyo.ac.jp 蓑島(21年度受講者)

es2.gi.k.u-tokyo.ac.jp 李(21年度受講者)

es3.gi.k.u-tokyo.ac.jp 林(21年度受講者)

es4.gi.k.u-tokyo.ac.jp 福田(21年度受講者)

es5.gi.k.u-tokyo.ac.jp 募集中

es6.gi.k.u-tokyo.ac.jp 仲里猛留(20年度受講者)

es7.gi.k.u-tokyo.ac.jp 中谷洋一郎(講師)

es8.gi.k.u-tokyo.ac.jp 劉暄暄(20年度受講者)

es9.gi.k.u-tokyo.ac.jp 宗永雅樹(20年度受講者)

es10.gi.k.u-tokyo.ac.jp 村中真人(20年度受講者)

es11.gi.k.u-tokyo.ac.jp 吳紅艷(20年度受講者)

es12.gi.k.u-tokyo.ac.jp 近藤修平(20年度受講者)

es13.gi.k.u-tokyo.ac.jp 白井和英(20年度受講者)

es14.gi.k.u-tokyo.ac.jp 中谷洋一郎(講師)

scmd.gi.k.u-tokyo.ac.jp 中谷洋一郎(講師)、齊藤太郎(講師補助)