

# 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。

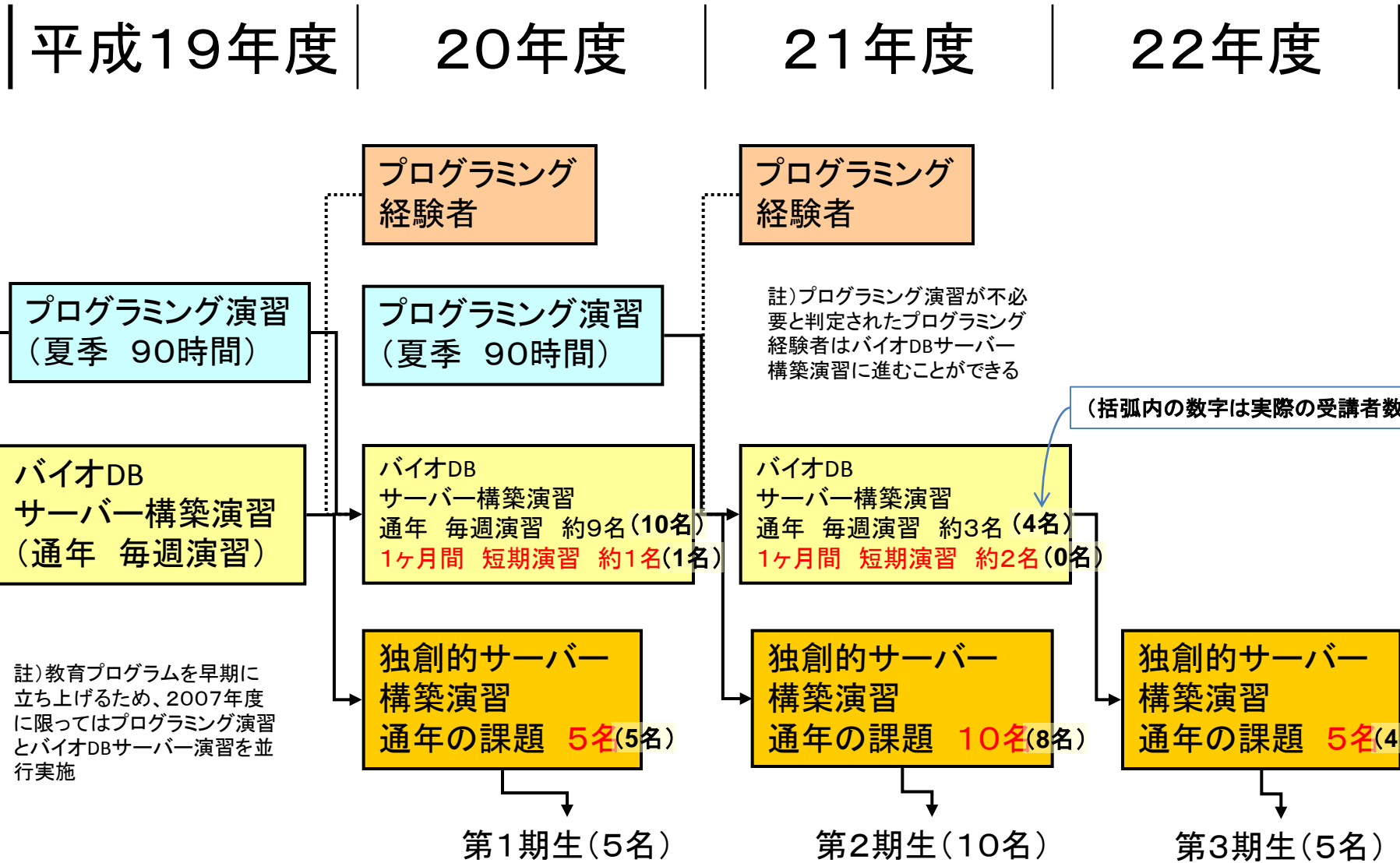
## ② プログラミング演習

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



演習用WS15台  
(平成19年度予算申請)

注) 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

# 受講者数

## バイオDBサーバー構築演習

### ● 平成19年度受講者

- 東大情報生命科学専攻から5名

### ● 平成20年度受講者

- 東大情報生命科学専攻から8名
- DBCLSから2名
- 自治医大から1名

### ● 平成21年度受講者

- 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

## 独創的サーバー構築演習

### 平成20年度受講者

- 東大情報生命科学専攻から5名

### 平成21年度受講者

- 東大情報生命科学専攻から6名
- DBCLSから2名

### 平成22年度受講者

- 東大情報生命科学専攻から4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を設置する
- 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

# バイオDBサーバー構築演習の概要 (平成21年度の例)

	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、セキュリティ設定、ネットワーク設定

- 演習受講者にサーバーを1台ずつ割り当て、OSインストールから演習する。
- OSインストール
  - 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。
- セキュリティ設定
  - 脆弱性について基礎的な事項を学習 (buffer overflow等)。
  - Yum-cronによる定期的なセキュリティアップデート。
- ネットワーク設定
  - 公開鍵認証方式によるssh接続の設定。
  - ファイアウォールの設定。



# Webサーバーの設置、Pukiwikiの設置、 シェルスクリプトプログラミング

- 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の設置し、ウェブ上で情報の共有と整理を多人数で行う。また演習ノートをPukiwikiで作成する。
  - 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。
- シェルスクリプトプログラミング
  - 文法の解説。
  - シェルスクリプトを使ったCGIプログラムの作成。

# Perlプログラミング、CGIプログラミング

- EnsemblのコードはPerlで書かれておりミラーサイト構築に必要となるため、Perlは時間をかけて詳しく解説した。
  - 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, ソート等。
  - 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。
  - CPAN (Comprehensive Perl Archive Network)によるソフトウェア・モジュールのインストールを解説。
    - BioPerlのインストール。
    - Makeによるインストールも解説。
- CGIプログラミング
  - HTTP, CGI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。

# データベースの設置、 Perlを使ったデータベース検索

- 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。
  - 遺伝子データ(TreeFam, Ensemblのデータを使用)をサーバーにダウンロードしMySQLで検索する。
- Perlを使ったデータベース検索CGIの作成。(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。)
  - PerlのDBIモジュールを使ってデータベースにアクセスし遺伝子データを検索。
  - BioPerlを使用した遺伝子系統樹解析。
  - CGI作成用のPerlモジュール HTML::Template, HTML::FillInformを使用し、データベースを検索し検索結果を画像で表示するCGIを作成。

# Java プログラミング演習

- 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

- Ensembl ミラーサイト構築。
  - yum, CPANを使って必要なモジュール・外部プログラムをインストール。
  - Ensembl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。
  - TeraStationをサーバーにマウントする。
  - “mysqldump”によるデータベースのバックアップとデータ復旧。
  - “rsync”コマンドによるバックアップとデータ復旧。

# バイオDBサーバー構築演習の進捗

平成19年度

平成20年度

平成21年度

- |   |   |   |                                |
|---|---|---|--------------------------------|
| 済 | 済 | 済 | 1. CentOS を自分のマシンにインストールする     |
| 済 | 済 | 済 | 2. ネットワークと接続する                 |
| 済 | 済 | 済 | 3. セキュリティアップデートを行う             |
| 済 | 済 | 済 | 4. web サーバーを立てる                |
| 済 | 済 | 済 | 5. CGIを設置する                    |
| 済 | 済 | 済 | 6. MySQL サーバーを立てる              |
| 済 | 済 | 済 | 7. 簡単なデータベース作成をする              |
| 済 | 済 | 済 | 8. Ensembl core をインストールしミラーを作成 |
|   |   |   | 9. Ensemblの完全ミラーを作る            |
|   |   | 済 | 10. バックアップを作成する                |

# 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。
  - バイオデータの解析演習。
  - 大規模並列計算演習。
    - Sun Grid Engineを用いた並列計算。
  - ウェブアプリケーション開発演習。
    - ゲノムブラウザーによる解析データの視覚化。
  - 受講者の研究結果をデータベースとして実装・公開することが目標。

# バイオデータの解析演習

受講者が実際の研究で使用するツール・データの解説と簡単な演習を行った。

- 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。
- JFreeChartを使用したグラフの描画。
- 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)。



# Sun Grid Engineを用いた並列計算

次世代シーケンサーの大量データを解析するために、SGEを使った並列計算を習得することが目標。

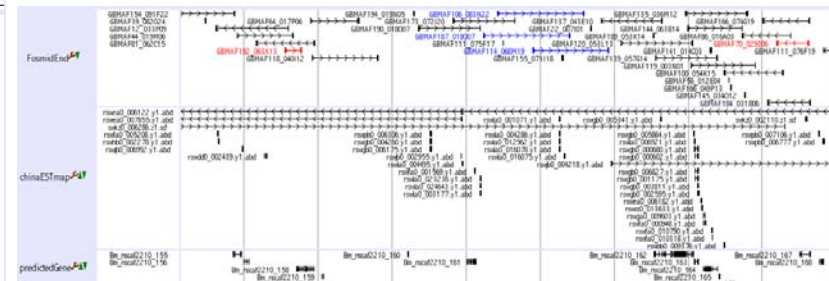
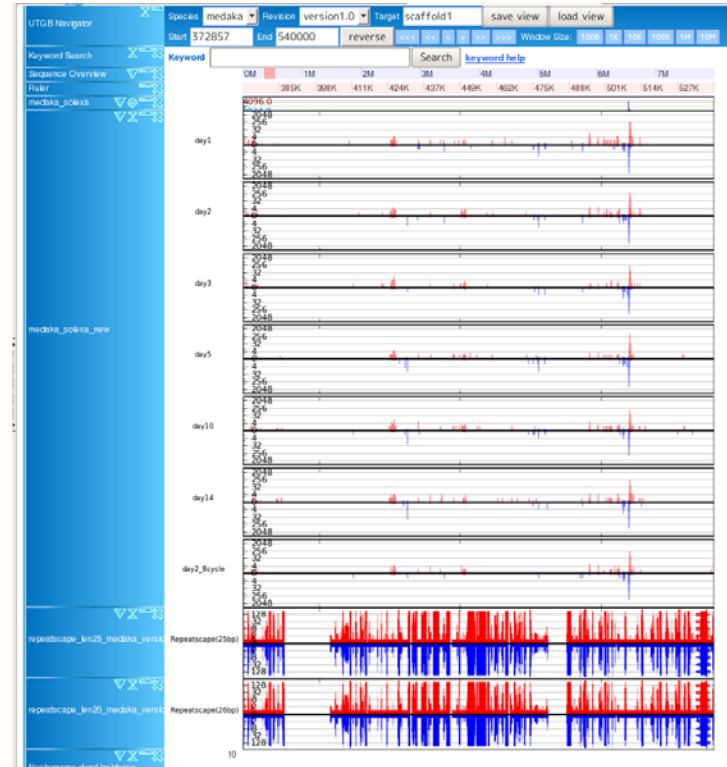
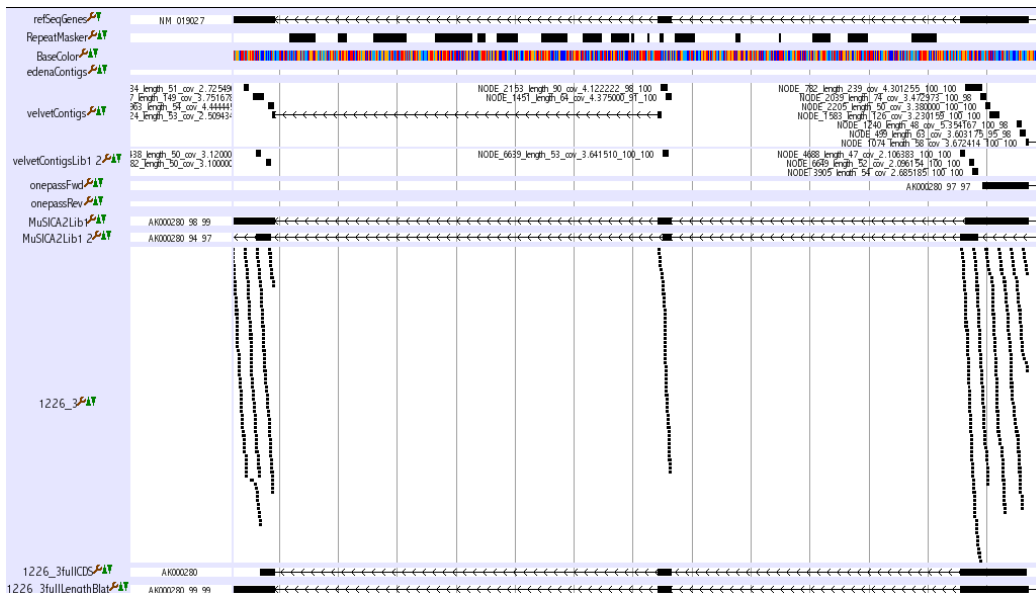
- SGEの使用法を解説
  - 並列ジョブの投入。
  - 実行/エラー状態の確認。
  - ジョブのキャンセル。
  - ジョブ間の依存関係の指定。
  - 必要なリソースを指定して実行。
  - キューの作成、エラー処理。
- SGEを使って、次世代シーケンサーで読んだリード配列をゲノムにマッピング。

# ウェブアプリケーション開発演習

- 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。
  - GWT(Google Web Toolkit)によるwebアプリケーション開発。
  - ユーザーインターフェイスの作成。
  - AJAXによるデータ通信。
  - JDBCによるデータベース検索。
  - 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。

# データベースの実装・公開

- 受講者が研究で使用する新規データをゲノムブラウザに表示する。
  - 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。

Online Mapping

Sequence

```
>no_name
ACGGGAAGAAAACAAAACCTTAATGGAAAAAGTAAACAAGCAACAGCAAAACGTTGGCCAAAGA
CAGCAAAATATCACTACAGCAATGTACAGCATTGAAGTACCAATAAATACATCCCATTTTA
TTCTGAACTCAAGTATTTCTGAGTCCCAGTTAAACAAATGTTCCCTTTTCAGCCCAA
TTACACCTGTCTGTTTCACTTTTGTCCCTTGACACGGCGAGCAAAACCGTGGCCGTCGACC
CGTGTGACAGCAACTAGAACAACACTTGTATGAGACTGAGGAGATGGGGTTGTGAGGAGA
ACCCATCTGGGTGAGAACCTTATCCAGCCATTGCAACGGGCCATGCAGGTGCAC TTC AAT
CCAGCAGGGGGTGTGGTCAACATCCTGACGGTGGTATTTCAGCCCCCATCCCTTGACGAA
GCTCATGGGATGGTGCACATCTTGGTGAAGTTCGATACACCACCTCGAAAGCCGTGGTTGAC
CGACTGGGCGAGGAGCTGGGCGAACAGCTGGTTGTTGAAGATCTTGGAGGTGCATCCGCT
GGGGATCTTGCACACTGTTGGTGGGGTGGAAAGCCATGCTGGAAATTGCAGTTGCGGCTTTG
GACAAAGATGCTGCTGTCGCTCAGACACTCTGCGTACACCTCCCGCCACAGTAGTACAG
```

Species: Medaka 1.0

Search Reset

Paste in your query sequence to find its location on the genomic sequences of specified species and revision. The online mapping system returns the locations found by BLAT alignment. The system accepts nucleotide sequences in the FASTA format or one flat string as input. Only sequences of length 18 - 100,000 bases will be processed.

ID: Search

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  
ACGGGAAGAAAACAAAACCTTAATGGAAAAAGTAAACAAGCAACAGCAAAACGTTGGCCAAAGA  
|||||  
ACGGGAAGAAAACAAAACCTTAATGGAAAAAGTAAACAAGCAACAGCAAAACGTTGGCCAAAGA  
CAGCAAAATATCACTACAGCAATGTACAGCATTGAAGTACCAATAAATACATCCCATTTTA  
|||||  
CAGCAAAATATCACTACAGCAATGTACAGCATTGAAGTACCAATAAATACATCCCATTTTA  
TTCTGAACTCAAGTATTTCTGAGTCCCAGTTAAACAAATGTTCCCTTTTCAGCCCAA  
|||||  
TTCTGAACTCAAGTATTTCTGAGTCCCAGTTAAACAAATGTTCCCTTTTCAGCCCAA  
TTACACCTGTCTGTTTCACTTTTGTCCCTTGACACGGCGAGCAAAACCGTGGCCGTCGACC  
|||||  
TTACACCTGTCTGTTTCACTTTTGTCCCTTGACACGGCGAGCAAAACCGTGGCCGTCGACC  
CGTGTGACAGCAACTAGAACAACACTTGTATGAGACTGAGGAGATGGGGTTGTGAGGAGA  
|||||  
CGTGTGACAGCAACTAGAACAACACTTGTATGAGACTGAGGAGATGGGGTTGTGAGGAGA  
ACCCATCTGGGTGAGAACCTTATCCAGCCATTGCAACGGGCCATGCAGGTGCAC TTC AAT  
|||||  
ACCCATCTGGGTGAGAACCTTATCCAGCCATTGCAACGGGCCATGCAGGTGCAC TTC AAT  
CCAGCAGGGGGTGTGGTCAACATCCTGACGGTGGTATTTCAGCCCCCATCCCTTGACGAA  
|||||  
CCAGCAGGGGGTGTGGTCAACATCCTGACGGTGGTATTTCAGCCCCCATCCCTTAAGG...  
-----CTTGACGAAGCTCATGGGATGGTGCACATCTTGGTGAAGTTCGTACACCACTTC  
|||||  
TGTGACTTGACGAAGCTCATGGGATGGTGCACATCTTGGTGAAGTTCGTACACCACTTC  
GAGCCCTGGTTGACCCACTGGGCGAGGAGCTGGGGCAACAGCTGGTTTGAAGATCTT  
|||||  
GAGCCCTGGTTGACCCACTGGGCGAGGAGCTGGGGCAACAGCTGGTTTGAAGATCTT  
GAGGCTGCATCCGCTGGGATCTTGCACACTGTGGTGGGTTGGAAGCCATCTGAAATT  
|||||  
GAGGCTGCATCCGCTGGGATCTTGCACACTGTGGTGGGTTGGAAGCCATCTGAAATT  
GCAGTTGCGGCTTTGGACAAAGATGCTGCTGCGCTAGACACTCTGGGTACACTCCCC  
|||||  
GCAGTTGCGGCTTTGGACAAAGATGCTGCTGCGCTAGACACTCTGGGTACACTCCCC  
GCCACGTAGTACAGGTGTAACC-----804-----CTTGCCATATGCTGCGCGT  
|||||  
GCCACGTAGTACAGGTGTAACCCTGGGA...CTTACCTTTGCTATGCTGCGCGT  
GTGCTCGATGGTGGAGTTGCGGTTGACCTTGGAAAGGAGGCCAGGCAAGAGCGTTCTT  
|||||  
GTGCTCGATGGTGGAGTTGCGGTTGACCTTGGAAAGGAGGCCAGGCAAGAGCGTTCTT  
GTGTTGACAGGGGTCAAGTGAAGCCGCTCCACCAAAGATGCTGTGG  
|||||  
GTGTTGACAGGGGTCAAGTGAAGCCGCTCCACCAAAGATGCTGTGG

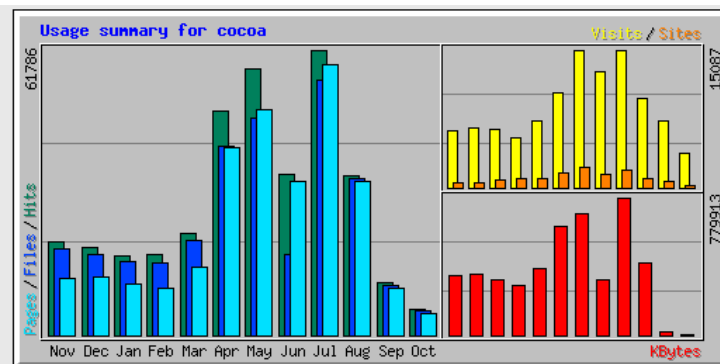
# 演習ノートをウェブ上で公開

過去の演習ノートの整理・改善を行った。

演習ノートのアクセス統計

The screenshot shows a Windows Internet Explorer browser window displaying the PukiWiki Plus website. The page has a clear navigation structure with several main sections:

- バイオDBサーバー構築演習**
- イントロダクション、OSのインストール**
  - イントロダクション
  - 最初の準備
  - イントロダクション、CentOSのインストール
  - CentOSのインストールに向けて
  - CentOSのインストールに向けて
  - VMware Server 上で CentOS をインストールする
- セキュリティ設定、ネットワーク設定**
  - セキュリティと定期アップデート
  - セキュリティと定期アップデート
  - セキュリティと定期アップデート、SSHによる外部からの接続
  - Linux とネットワークの基礎
  - ネットワーク設定、SSHによる外部からの接続
- webサーバーの設置、Pukiwikiの設置、シェルスクリプト**
  - web サーバーに動的なコンテンツを追加する
    - <http://www.xerial.org/maven/repository/site/xerial-project/docs/WebApplication.html>
  - web サーバーの設置、シェルスクリプト、Pukiwikiの設置
  - pukiwikiの設置
  - Pukiwiki による情報共有
  - シェルスクリプト
  - シェルスクリプト、Pukiwikiの設置
- Perlプログラミング**
  - Perl 演習1-2



Summary by Month										
Month	Daily Avg				Monthly Totals					
	Hits	Files	Pages	Visits	Sites	KBytes	Visits	Pages	Files	Hits
<a href="#">Oct 2010</a>	365	347	321	250	300	5829	3764	4820	5217	5482
<a href="#">Sep 2010</a>	379	357	337	246	630	16046	7380	10125	10730	11387
<a href="#">Aug 2010</a>	1114	1096	1073	315	978	406870	9778	33287	34006	34536
<a href="#">Jul 2010</a>	1993	1782	1891	486	1906	779913	15087	58623	55244	61786
<a href="#">Jun 2010</a>	1161	580	1108	420	1452	316922	12628	33258	17428	34858
<a href="#">May 2010</a>	1861	1515	1575	483	2230	689661	15002	48825	46988	57704
<a href="#">Apr 2010</a>	1622	1362	1354	344	1636	614311	10320	40644	40863	48669
<a href="#">Mar 2010</a>	715	658	479	235	1024	375322	7287	14878	20401	22188
<a href="#">Feb 2010</a>	627	562	364	194	956	286172	5440	10212	15742	17572
<a href="#">Jan 2010</a>	558	516	359	206	795	314117	6397	11144	16014	17301
<a href="#">Dec 2009</a>	610	567	412	210	589	343869	6512	12779	17606	18910
<a href="#">Nov 2009</a>	672	627	407	206	611	338368	6192	12217	18811	20181
<b>Totals</b>						<b>4487401</b>	<b>105787</b>	<b>290812</b>	<b>299050</b>	<b>350574</b>

# 受講者の進路

(卒業生の就職先、または現在の所属)

- DBCLS : 2名
- 研究員 : 2名
- 博士課程在学中 : 5名
- 修士課程在学中 : 3名
- 就職 : 6名
  - IBM, JR東日本, 日立, 野村総研など
- 不明、その他 : 3名