

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。

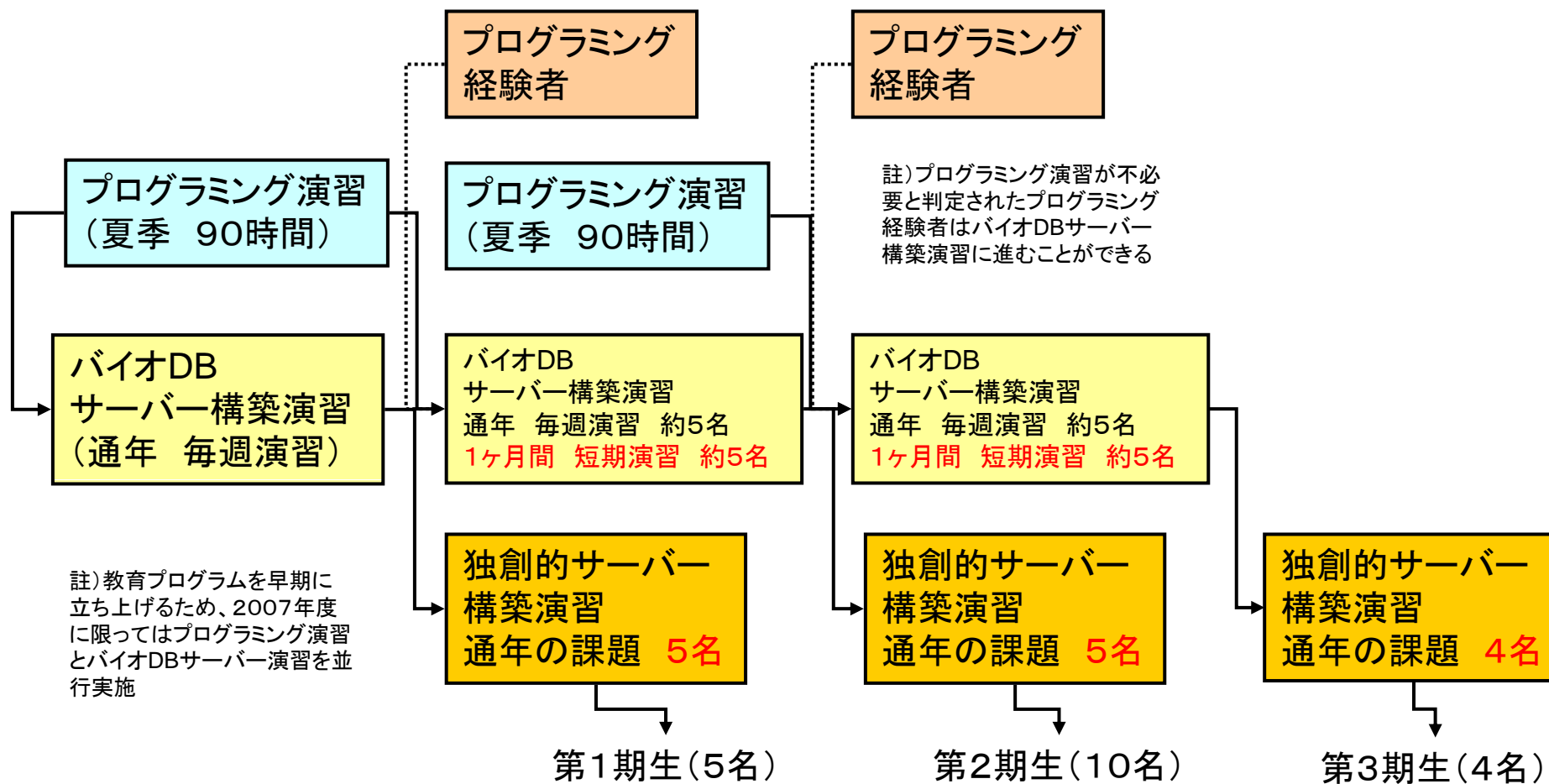
年次計画

平成19年度

20年度

21年度

22年度



演習用WS15台
(平成19年度予算申請)

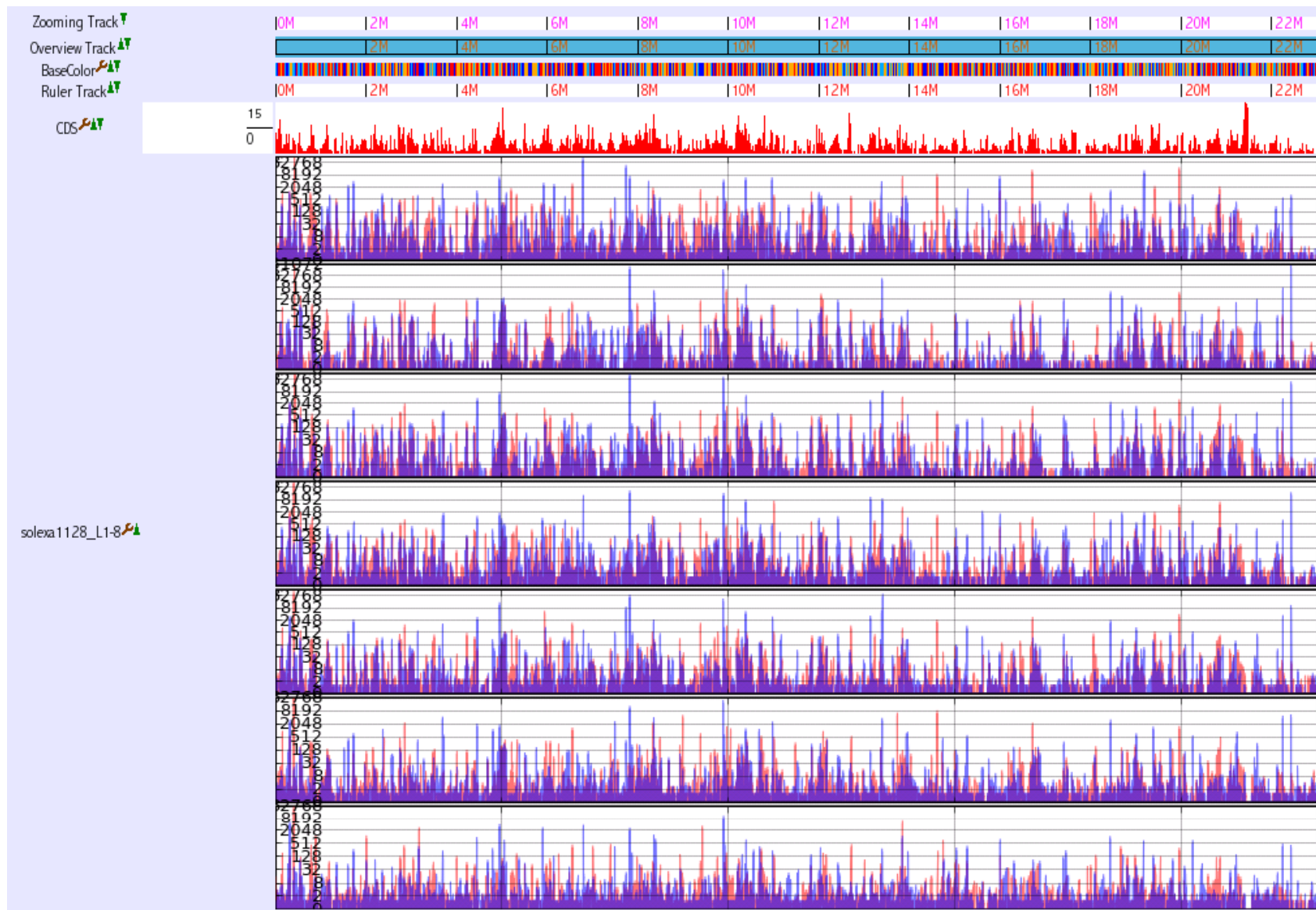
注) 1期生と2期生が20年度には重なること(21年度は2, 3期生)、WSが15台であること、演習スタッフ1.5名による徒弟制度であるため、各年15名の受け入れが限度である

平成22年度演習

- 受講者
 - 東大情報生命科学専攻から4名
- 演習目標
 - 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。
 - 統合DBプロジェクトにおいて次世代シーケンシング情報のデータベースを構築できる人材を育成する。
- 講義内容
 - Sun Grid Engineのインストール。
 - Sun Grid Engineを用いた並列計算。
 - 次世代シーケンサーのデータをゲノムにマッピング。
 - 次世代シーケンサーのデータをゲノムブラウザに表示する。

独創的サーバー構築演習 例

- 過去の受講者の演習例
 - ✓ 超高速シーケンサー Solexa の base call
 - ✓ 並列 BLAST / BLAT を使った短いタグ(25–36 nt)のアライメント
 - ✓ 全長 cDNA 推定アルゴリズムの工夫
 - ✓ 5' end タグを使った遺伝子発現量解析
 - ✓ 解析パイプラインの研究開発
 - ✓ 表示ルールの研究開発
 - ✓ 従来の遺伝子情報に比べて数百倍のデータ量を適切な応答時間で処理する工夫

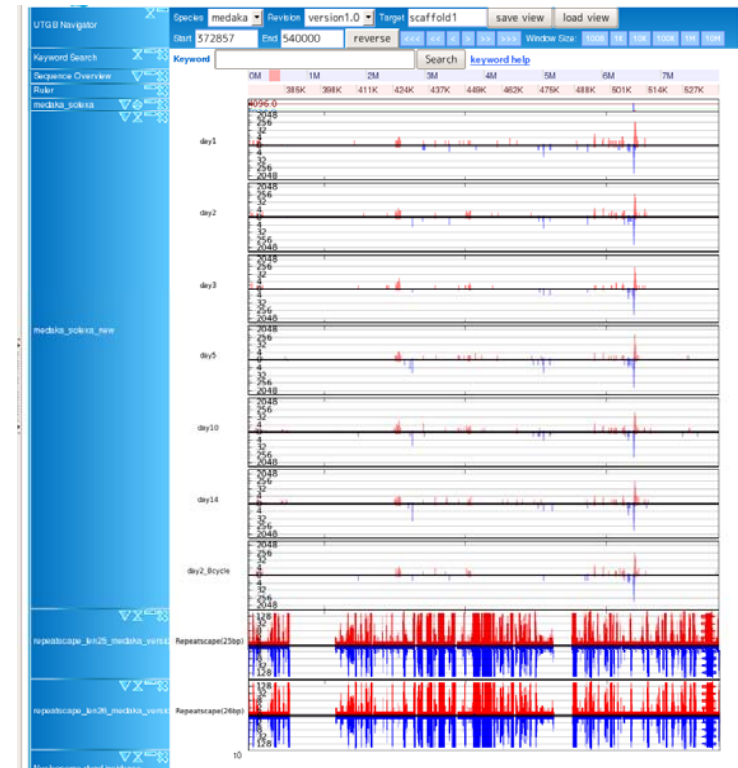
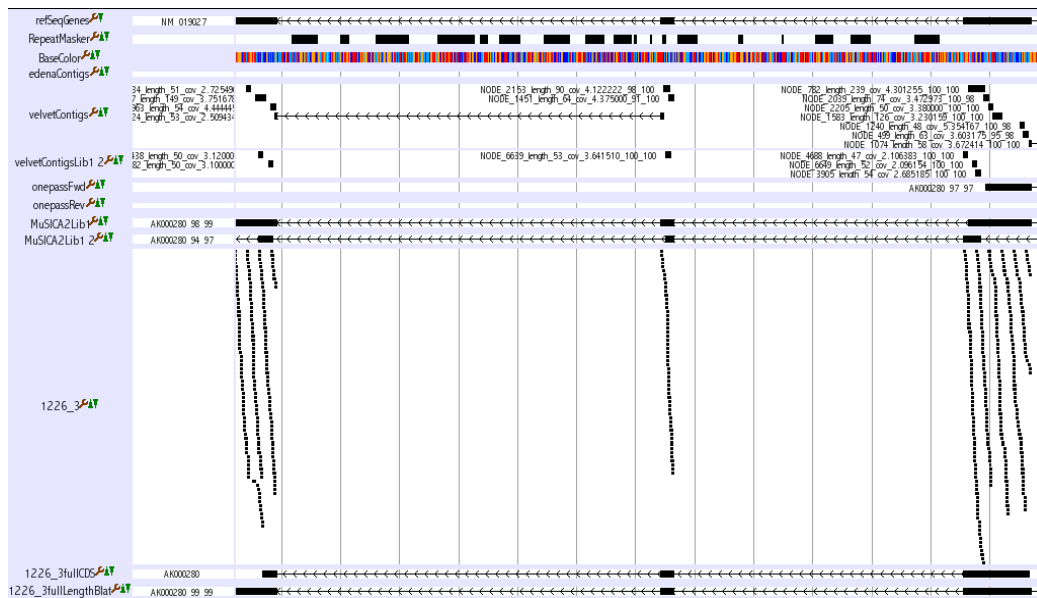


Solexaタグの表示例

従来の遺伝子情報に比べて数百倍のデータ量を適切な応答時間で処理する工夫

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

Online Mapping

Sequence

```
>no_name
ACGGGAAGAAAACAAAACTTAATGGAAAAAGTAACAAAGCAACAGCAAACGTTGGCCAAAGA
CAGCAAAATATCACTACAGCAATGTACAGCATTGAAAGTACCAATAAATACATCCCATTTTA
TTCTGAACCTCAAGTATTTCTGAGTCCCAGTTAACAAATGTTCCCTTCTTTTCAGCCCAA
TTACACCTGTCTGTTTCACCTTTGTCCCTTGACACGGCGAGCAAACCGTGGCCGTCGACC
CGTGTGACAGCAACTAGAACACACTTGTATTGAGACTGAGGAGATGGGGTTGTGAGGAGA
ACCCATCTGGGTGAGAACCTTATCCAGCCATTGCAACGGGCCATGCAGGTGCACCTCAAT
CCAGCAGGGGGTCTGGTACATCCTGACGGTGTATTTCAGCCCCCATCCCTTGACGAA
GCTCATGGCGATGGTGCACATCTTGGTGAAGTTCGTACACCACCTCGAAAGCGGTTGTGAC
CGACTGGCGGAGGAGCTGGCGAAGCAGCTGGTTGTTGAAAGATCTTGAGGCTGCATCCGCT
GGGGATCTTGACACTGTGGTGGGGTGGAAAGCCATGCTGGAAATTCAGTTGGCGCTTTG
GACAAAATGCTGCTGTCGCTCAGACACTCTGCGTACACCTCCCCGCCACGTAGTACAG
```

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. N's Q gap Q gap T gap T gap strand Q
count bases count bases count bases count bases nan
View 793 0 0 0 1 1 2 884 + no_
```

>no_name:0+794 of 794 scaffold1211:611860+613537 of 1085344

```
ACGGGAAGAAAACAAAACTTAATGGAAAAAGTAACAAAGCAACAGCAAACGTTGGCCAAAGA
ACGGGAAGAAAACAAAACTTAATGGAAAAAGTAACAAAGCAACAGCAAACGTTGGCCAAAGA
CAGCAAAATATCACTACAGCAATGTACAGCATTGAAAGTACCAATAAATACATCCCATTTTA
CAGCAAAATATCACTACAGCAATGTACAGCATTGAAAGTACCAATAAATACATCCCATTTTA
TTCTGAACCTCAAGTATTTCTGAGTCCCAGTTAACAAATGTTCCCTTCTTTTCAGCCCAA
TTCTGAACCTCAAGTATTTCTGAGTCCCAGTTAACAAATGTTCCCTTCTTTTCAGCCCAA
TTACACCTGTCTGTTTCACCTTTGTCCCTTGACACGGCGAGCAAACCGTGGCCGTCGACC
TTACACCTGTCTGTTTCACCTTTGTCCCTTGACACGGCGAGCAAACCGTGGCCGTCGACC
CGTGTGACAGCAACTAGAACACACTTGTATTGAGACTGAGGAGATGGGGTTGTGAGGAGA
CGTGTGACAGCAACTAGAACACACTTGTATTGAGACTGAGGAGATGGGGTTGTGAGGAGA
ACCCATCTGGGTGAGAACCTTATCCAGCCATTGCAACGGGCCATGCAGGTGCACCTCAAT
ACCCATCTGGGTGAGAACCTTATCCAGCCATTGCAACGGGCCATGCAGGTGCACCTCAAT
CCAGCAGGGGGTCTGGTACATCCTGACGGTGTATTTCAGCCCCCATCC-----80-
-----CTTGACGAAGCTCATGCGGATGGTGCACATCTTGGTGAAGTTCGTACACCACTC
TTGTTACCTTGACGAAGCTCATGCGGATGGTGCACATCTTGGTGAAGTTCGTACACCACTC
GAAGCCGTTGTTGACCGACTGGCGAGGAGCTGGCGGAACAGCTGGTTGTTGAAGATCTT
GAAGCCGTTGTTGACCGACTGGCGAGGAGCTGGCGGAACAGCTGGTTGTTGAAGATCTT
GAGGCTGCATCCGCTGGGATCTTGCACACTGTGGTGGGGTGGAGCCATGCTGGAAAT
GGGCTGCATCCGCTGGGATCTTGCACACTGTGGTGGGGTGGAGCCATGCTGGAAAT
GCAGTTGCGGCTTTGGACAAAGATGCTGCTGCCTGCAGACACTTGCCTACACCTCCCC
GCAGTTGCGGCTTTGGACAAAGATGCTGCTGCCTGCAGACACTTGCCTACACCTCCCC
GCCACCTAGTACAGGTGAACC-----804-----CTTGCCTATGTGCTGCGCGCT
GCCCCAGTAGTACAGGTGAACCTCGGGA...CTCTACCTTTGCTATGTGCTGCGCGT
GTGCTGCATGGTGGAGTTGCGGTTGACGTTGGAAGGAGGCCAGCCAGGAGAAGCGGTTCTT
GTGCTGCATGGTGGAGTTGCGGTTGACGTTGGAAGGAGGCCAGCCAGGAGAAGCGGTTCTT
GTTGTTGCAGGGGTCAGTGAAGCCGTCACCAAAGATGCTGTGG
GTTGTTGCAGGGGTCAGTGAAGCCGTCACCAA--GATGCTGTGG
```

演習ノート (毎週の講義内容と宿題集)



SeminarScribe
<http://mlab.cb.k.u-tokyo.ac.jp/~mkasa/ensemblmirror/index.php?SeminarScribe>

[ホーム | 一覧 | 単語検索 | 最終更新 | ヘルプ] [新規 | 編集 | 添付] [no Trackback]

最新の20件

2007-12-17
・ [演習ノート/20071011](#)
・ [演習ノート/20070913](#)

2007-12-14
・ [統合データベース支援: DB構築者の養成におけるバイオDBサーバー構築演習](#)

2007-12-13
・ [WebAppDevelopmentSchedule](#)

2007-12-07
・ [演習ノート/20070516その2](#)

2007-11-08
・ [演習ノート/20070705](#)
・ [ソース課題1](#)

2007-11-04
・ [演習ノート/20070516その3](#)
・ [Ensemblインストールのメモ](#)

2007-11-02
・ [演習ノート/20070418](#)

2007-11-01
・ [GWTUserInterface](#)

2007-10-30
・ [演習ノート/20070802](#)
・ [SeminarScribe](#)
・ [演習ノート/20070412](#)

2007-10-25
・ [EnsemblMirror](#)
・ [SeminarPowerpoints](#)

2007-10-24
・ [演習ノート/20070627](#)

2007-10-23
・ [演習ノート/20070425](#)
・ [演習ノート/20070530](#)

2007-10-08
・ [演習ノート/20071004](#)

Top > SeminarScribe

- ・ [演習ノート](#)
 - [ノート一覧](#)

演習ノート

演習ノートはこのページからリンクしてください。下記一覧に自分の担当の演習ノートへのリンクが無い場合は、同じような書式で追加してください。

ノート一覧

- ・ [4/6 インタロダクション](#)
- ・ [4/9 最初の準備](#)
- ・ [4/12 CentOSのインストールに向けて](#)
- ・ [4/18 Linux とネットワークの基礎](#)
- ・ [4/25 VMware Server 上で CentOS をインストールする](#)
- ・ [5/9 CentOS 上でweb サーバーを設置する](#)
- ・ [5/16 web サーバーに動的なコンテンツを追加する](#)
- ・ [5/16 その2 pukiwikiの設置](#)
- ・ [5/16 その3 シェルスクリプト](#)
- ・ [5/23 セキュリティと定期アップデート](#)
- ・ [5/25 Pukiwiki による情報共有](#)
- ・ [5/30 RDBMS を使ってみる](#)
- ・ [6/6-13 Perl 演習1-2](#)
- ・ [6/27 Perl 演習3](#)
- ・ [7/05 PerlでCG演習](#)
- ・ [7/19 tarballからソフトのインストールをする](#)
- ・ [8/2 CPANを使いこなす](#)
- ・ [9/13 Ensemble core](#)
- ・ [10/4 ネットワークトラブルへの対処](#)
- ・ [10/11いろいろ](#)

Link: [演習ノート/20071011\(8d\)](#) [演習ノート/20070913\(8d\)](#) [演習ノート/20070516その2\(17d\)](#) [演習ノート/20070705\(46d\)](#) [演習ノート/20070516その3\(50d\)](#) [演習ノート/20070418\(52d\)](#) [演習ノート/20070802\(55d\)](#) [演習ノート/20070412\(55d\)](#) [EnsemblMirror\(60d\)](#) [演習ノート/20070627\(62d\)](#) [演習ノート/20070425\(62d\)](#) [演習ノート/20070530\(63d\)](#) [演習ノート/20071004\(78d\)](#) [演習ノート/20070525\(80d\)](#) [演習ノート/20070516\(81d\)](#) [演習ノート/20070719\(90d\)](#) [演習ノート/20070523\(103d\)](#) [演習ノート/20070509\(104d\)](#) [演習ノート/20070606\(104d\)](#) [演習ノート/20070409\(157d\)](#) [演習ノート/20070406\(167d\)](#)

<http://mlab.cb.k.u-tokyo.ac.jp/~mkasa/ensemblmirror/index.php>