グローバルCOEプログラム系統講義「ベーシックサイエンスコース」

バイオインフォマティクス：
生物系研究者における情報リテラシーとしての配列解析

Bioinformatics: Sequence analyses as information literacy of life scientists

嶋田　誠
（藤田保健衛生大学 遺伝子発現機構学）

1日目：2010年 9月28日（火） 17:00−18:30　基礎医学研究棟1階会議室
2日目：2010年10月 4日（月） 18:00−19:30　基礎医学研究棟1階会議室

アウトライン

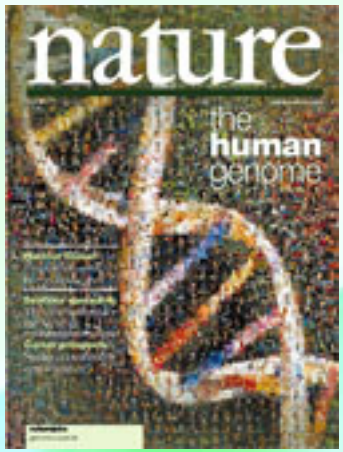
- [第1回]
  - 概論
    - ねらい
    - バイオインフォマティクス：どこまで把握すべきか？
  - データベース
- [第2回]
  - ゲノム関連大規模プロジェクト
  - バイオインフォマティクス独学のコツ
  - 生命科学者（実験研究者）にとってのプログラミング
    - アルゴリズム
  - データの形式と変換
    - データ加工変換
    - パース
  - ツール群の分類と見つけ方
    - 配列研究関連ツールの紹介

生物情報に影響を与えた大規模プロジェクト

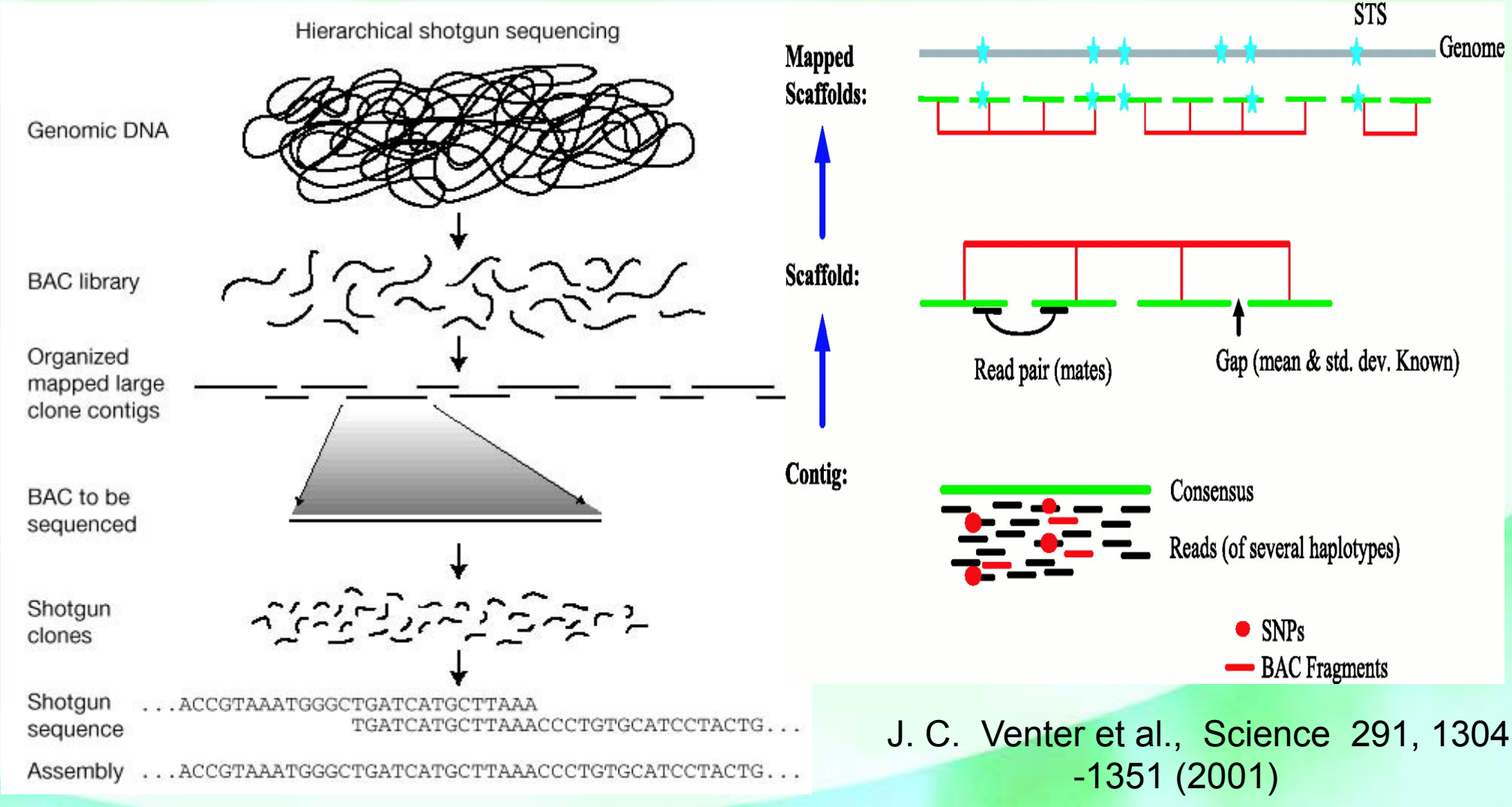
- Human Genome
- cDNA
  - Full-length cDNA Japan
  - H-Invitational
  - FANTOM
- ENCODE
- HapMap
- 1000 Genomes

ゲノムプロジェクト

- 1987 和田昭充、機械による自動配列決定戦略提案
- 1990 アメリカ、イギリス、日本、ドイツ、フランスを中心とする国際コンソーシアム(INC)合意
  - 統一戦略なし、map作成(YAC,BAC,PAC)、新技術開発、
- 1996 バミューダ会議
  - ルール(参加、実施前公表、分担、2005年完成目標)
- 1998 Celera社3年間でヒトゲノム概要版解読宣言
  - [INC:目標変更、BAC概要版データは即座に公共DBにて公開]
- 2000 Jun. 概要版完成宣言
  - [INC] Nature (2/15/2001)
  - [Celera] Science (2/16/2001)
- 2003 Apr. [INC]全染色体解読完了宣言
  - [INC] Nature (10/21/2004)



Assembly方法：
clone-based 階層的vs. random whole genome shotgun sequencing strategies (WGS)



J. C. Venter et al., Science 291, 1304 -1351 (2001)

INC, Nature 409, 860-92 (2001)

結果の利用：階層的 vs. 全ゲノム

- ハプロタイプ由来 vs. ５人のコンセンサス
  - [INC] BACクローンのそれぞれが単一ハプロタイプ(ただし多くの不特定個人から作成されたBACライブラリー)
  - [Cerela] ５人の個人由来（ただしJ. Craig Venter氏由来が最も多い）のBAC配列の重なりを頼りに作られたコンセンサス配列

cDNA projects

- EST project
  - J. Craig Venter (USA)が先駆け
  - dbEST(NCBI)の統計ではヒトが83億エントリで最大。
- Full length project
  - H-Inv
  - FANTOM

Human Full-Length cDNA Annotation Invitational (http://hinv.jp)

•ヒト完全長cDNAプロジェクト日本(FLJ)で日本は世界をリードするコレクションを有していた(eg., Nature Genetics 36, 40 - 45, 2004)。
•H-Invitational(お台場マラソン) 2002年8月25日-9月3日

項目数	提供機関
2,031	Kazusa DNA Research Institute
397	Full Length cDNA Japan / Kazusa DNA Research Institute
6,374	Full Length cDNA Japan / 東京大学医科学研究所
22,047	Full Length cDNA Japan / Helix Research Institute, Inc.
9,212	German Human cDNA Project （DKFZ、ドイツ）
15,600	Mammalian Gene Collection （NCI/NIH、アメリカ）
758	Human cDNAs （Chinese National Human Genome Center、中国）
56,419	Total

•ヒト完全長cDNAにヒト全mRNAのデータを加えて、公開した。(2006)
- 現在でも、完全長データに絞った検索が可能
- (eg., AS subDB, H-DBAS, http://hinv.jp/h-dbas/)

FANTOM

http://www.osc.riken.jp/contents/fantom/

- FANTOMは、理化学研究所のマウスエンサイクロペディアプロジェクトで収集された完全長cDNAのアノテーションを目的とした国際研究コンソーシアム
- 2000年に結成され、2007年12月現在、参加国数15カ国、参加機関は国内外合わせて51機関
- FANTOM データベース　http://fantom.gsc.riken.jp/
- 研究ツール
  - Genome Browser
  - EdgeExpressDB：制御関係と遺伝子やプロモーター活性との関連を表示
  - SwissRegulation：モチーフ活性の応答解析
- 完全長cDNA クローンバンク

ENCODE (Encyclopedia of DNA Elements) project (1), http://www.genome.gov/ENCODE/

- launched by the National Human Genome Research Institute (NHGRI) in 2003
- goal: defining the functional elements in the human genome (protein, RNA, regulatory elements, etc.)
- Pilot phase: Nature 447, 799-816 (2007)
  - Focusing on 1% (44 regions) of the human genome
    - pervasive transcription of genome
    - transcription start sites (regulatory sequences, chromatin accessibility and histone modification)
    - chromatin structure; relation with DNA repication, transcriptional regulation
    - evolutionary insights concerning the functional landscape of the human genome

ENCODE (Encyclopedia of DNA Elements) project (2), http://www.genome.gov/ENCODE/

- Two expansion (since 2007) to
  - the whole human genome
  - model organism ENCODE (modENCODE)
    - comprehensive annotation ot the functional elements in
      - the *C. elegans* and *D. melanogaster* genomes
- Browsers
  - http://genome.ucsc.edu/ENCODE/
  - http://www.sanger.ac.uk/PostGenomics/encode/
- Selected ENCODE Publications
  - Nucleic Acids Res. 2010 Jan;38:D620-5.
  - Nature. 2007 Jun 14;447(7146):799-816.
  - ENCODE Special Issue. Genome Research, 2007

HapMap project

HapMap Homepage | HapMap: Data

hapmap.ncbi.nlm.nih.gov/index.html.en

phase 19 / 20

International HapMap Project

Home | About the Project | Data | Publications | Tutorial

中文 | English | Français | 日本語 | Yoruba

The International HapMap Project is a partnership of scientists and funding agencies from Canada, China, Japan, Nigeria, the United Kingdom and the United States to develop a public resource that will help researchers find genes associated with human disease and response to pharmaceuticals. See "About the International HapMap Project" for more information.

Project Information

About the Project
HapMap Publications
HapMap Tutorial
HapMap Mailing List
HapMap Project Participants

Project Data

HapMap Genome Browser release #28 (Phases 1, 2 & 3 - merged genotypes & frequencies)
HapMap3 Genome Browser release #3 (Phase 3 - genotypes & frequencies)
HapMap Genome Browser release #27 (Phase 1, 2 & 3 - merged genotypes & frequencies)
HapMap3 Genome Browser release #2 (Phase 3 - genotypes, frequencies & LD)
HapMap Genome Browser release#24 (Phase 1 & 2 - full dataset)
GWAs Karyogram
HapMart
HapMap FTP
Bulk Data Download
Data Freezes for Publication
ENCODE Project
Guidelines For Data Use

Useful Links

TSC SNP Downloads
HapMap Samples at Coriell Institute
HapMap Project Press Release
NHGRI HapMap Page
NCBI Variation Database (dbSNP)
Japanese SNP Database (JSNP)

News

- 2010-08-18: **HapMap Public Release #28**

  Genotypes and frequency data in hapmap format are now available for data in merged HapMap phases I+II+III release #28 (NCBI build 36, dbSNP b126). Data is **available for bulk download** and also **available for browsing**. Click here to read the latest **release notes**.

- 2010-05-28: **HapMap3 Public Release #3**

  Genotypes and frequency data in hapmap format are now available for data in HapMap phase 3 release #3 (NCBI build 36, dbSNP b126). Data is **available for bulk download** and also **available for browsing**. Click here to read the latest **release notes**.

- 2010-05-28: **HapMap3 CNV Genotypes**

  Copy Number Variation genotypes for HapMap phase samples are **available for bulk download**.

- 2009-12-10: **Corrected HapMap3 phased haplotypes available for chromosome X**

  Phased haplotypes for consensus HapMap3 release 2 data for chromosome X has been corrected and the new data are now **available for bulk download**. Sorry for any inconvenience this might have caused.

- 2009-12-02: **HapMap3 phased haplotypes available for chromosome X**

  Phased haplotypes for consensus HapMap3 release 2 data has been phased for chromosome X and are now available for bulk download. [Update: The downloading was disabled because several users have found that there are repeating data in some of the chrX phasing data files. The data source is being contacted and the downloading will be enabled as soon as the problem is cleared.]

- 2009-11-18: **Short downtime for hardware maintenance**

  There will be a 30-minute downtime for Hapmap site on Nov. 23, 2009 between 8:30 am - 9:00 am (EST). Sorry for any inconvenience. [Update: Maintenance completed.]

- 2009-11-05: **NCBI Scheduled maintenance**

  NCBI servers will undergo maintenance beginning November 13 at 3:00 p.m. to November 14 at 8:00 p.m. (EST). Therefore, some NCBI services may be intermittently slow or inaccessible. Please contact NCBI with concerns: **info at ncbi.nlm.nih.gov.** [Update: NCBI Scheduled maintenance compleleted.]

- 2009-11-03: **Database maintenance notice**

  HapMap is scheduled to have a database maintenance from 5:00pm 11/06/2009 - 8:00am 11/09/2009 EST. The site will be inaccessible during the maintainence. Sorry for the inconvenience. [Update: Database maintenance compleleted.]

- 2009-04-02: **HapMap3 CEL files available**

  Raw signal intensity data from HapMap3 genotypes on the Genome-Wide Human SNP Array 6.0 are now **available for bulk download**.

- 2009-02-09: **HapMap3 Phased Haplotypes available**

参加組織

参加施設

Baylor College of Medicine (USA)
Beijing Genomics Institute (China)
Beijing Normal University (China)
Broad Institute of Harvard and MIT (USA)
Center for Statistical Genetics, University of Michigan (USA)
Chinese National Human Genome Center at Beijing (China)
Chinese National Human Genome Center at Shanghai (China)
Cold Spring Harbor Laboratory (USA)
Eubios Ethics Institute (Japan)
北海道医療大学 (Japan)
Hong Kong University of Science and Technology (China)
Howard University (USA)
Illumina (USA)
Johns Hopkins School of Medicine (USA)
McGill University & Génome Québec Innovation Centre (Canada)
ParAllele BioScience (USA)
Perlegen Sciences (USA)
理化学研究所 (Japan)
The Chinese University of Hong Kong (China)
The University of Hong Kong (China)
University of California, San Francisco (USA)
University of Ibadan (Nigeria)
University of Oxford (UK)
University of Oxford / Wellcome Trust Centre for Human Genetics (UK)
東京大学 (Japan)
University of Utah (USA)
Washington University, St. Louis (USA)
Wellcome Trust Sanger Institute (UK)

政府，財団等

Chinese Academy of Sciences
Chinese Ministry of Science and Technology
Delores Dore Eccles Foundation
Genome Canada
Génome Québec
Hong Kong Innovation and Technology Commission
文部科学省
Natural Science Foundation of China
The SNP Consortium
U.S. National Institutes of Health (NIH)
University Grants Committee of Hong Kong
Wellcome Trust
W.M. Keck Foundation

Last updated : index.html.ja 554 2009-10-01 19:55:33Z zhahua fc

Home | プロジェクトについて | データ | 文献 | Tutorial

このウェブサイトに関する質問やコメントは hapmap-help@ncbi.nlm.nih.govまでお願いします.

HapMap 特徴
http://hapmap.ncbi.nlm.nih.gov/

- 継続されているprojectである。
  - **最新**（2010-08-18）：phase III, release 28
- “phase”という語に注意
  - phases I II III vs. phased haplotype
- [応用]ヒトのLDとその集団間差異を明らかにした。
- データのダウンロード、および、
- 多型に関する、様々なブラウザ表示およびツール群をweb通じ提供している。

HapMapのおもな成果

- Phase I / IIでは、4集団270個体について、ゲノムワイドに約3.6M個のSNPsがタイピングされた。同時に、用いたサンプル中の全多型は10M個と推定された。LD領域マップの作成とヒトの疾患に関連する数百遺伝子座を同定できた(Nature 449, 851-861, 2007)。
- Phase IIIでは、7集団を加えた計1184個体で
  - 約1.6M個のSNPsをタイピング、
  - ENCODE領域の配列決定、
  - 約1.6K箇所のCNP領域の同定、
- を行い、稀な変異の集団間差異とその解析方法を示した(Nature 467, 52–58, 2010)。

Samples of HapMap3

label	population sample	number of samples	QC samples
ASW	African ancestry in Southwest USA	90	83
CEU	Utah residents with Northern and Western European ancestry from the CEPH collection	180	165
CHB	Han Chinese in Beijing, China	90	84
CHD	Chinese in Metropolitan Denver, Colorado	100	85
GIH	Gujarati Indians in Houston, Texas	100	88
JPT	Japanese in Tokyo, Japan	91	86
LWK	Luhya in Webuye, Kenya	100	90
MEX	Mexican ancestry in Los Angeles, California	90	77
MKK	Maasai in Kinyawa, Kenya	180	171
TSI	Toscans in Italy	100	88
YRI	Yoruba in Ibadan, Nigeria	180	167
Total		1301	1184

1000 Genomes(千人ゲノム)

http://www.1000genomes.org/

- 多数の個人ゲノム解読により、ヒト集団の多様性を明らかにする。
- [応用]GWAS
- Pilot Project

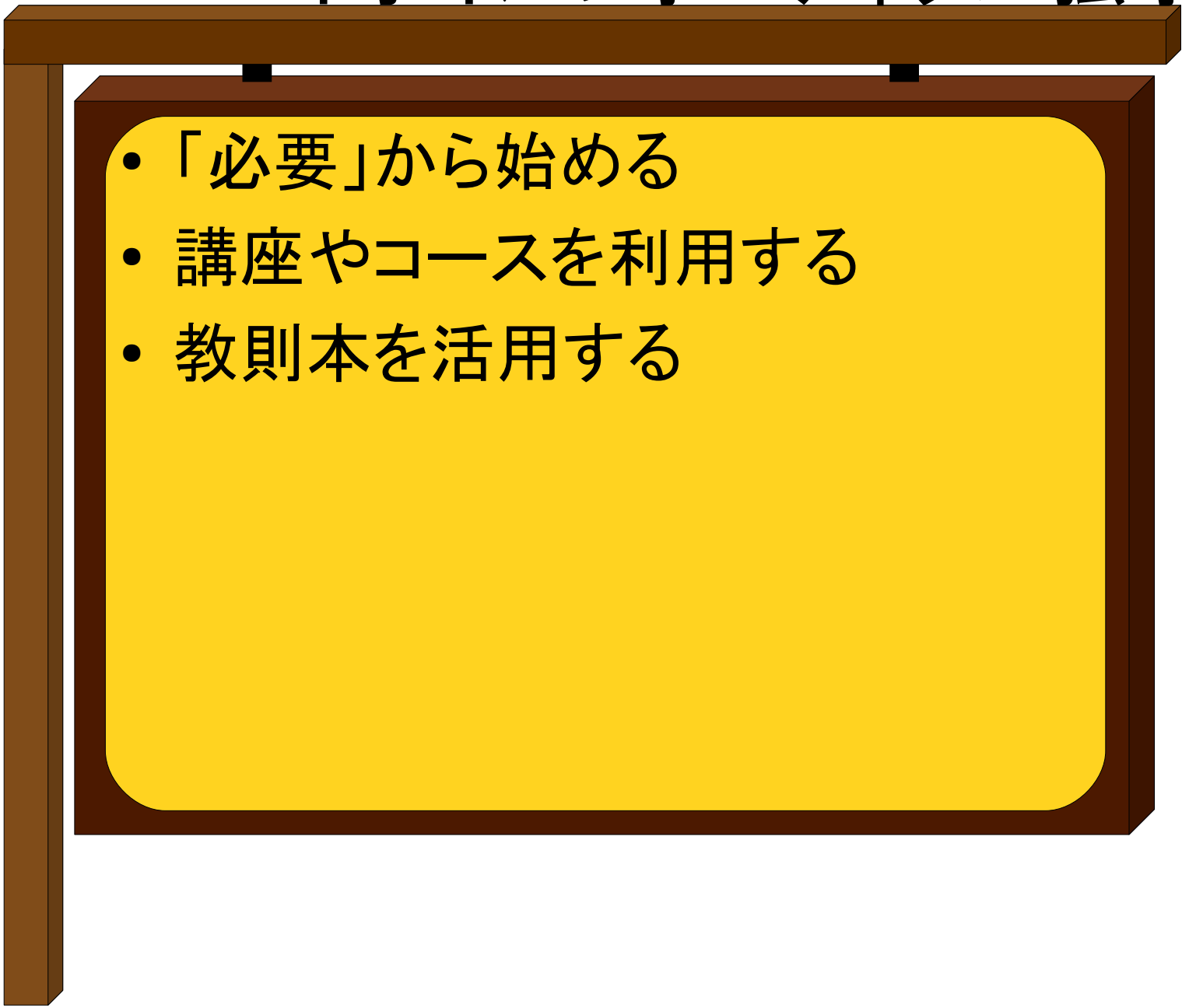
Pilot	Purpose	Coverage	Strategy	Status
1 - low coverage	Assess strategy of sharing data across samples	2-4X	Whole-genome sequencing of 180 samples	Sequencing completed October 2008
2 - trios	Assess coverage and platforms and centers	20-60X	Whole-genome sequencing of 2 mother-father-adult child trios	Sequencing completed October 2008
3 - gene regions	Assess methods for gene-region-capture	50X	1000 gene regions in 900 samples	Sequencing completed June 2009

- Main Project
  - 2000 samples at 4X
    - → 1st set: 1101 samples from 12 populations
    - → 2nd set: 899 samples from 10 populations

アウトライン

- [第1回]
  - 概論
    - ねらい
    - バイオインフォマティクス：どこまで把握すべきか？
  - データベース
- [第2回]
  - ゲノム関連大規模プロジェクト
  - バイオインフォマティクス独学のコツ
  - 生命科学者（実験研究者）にとってのプログラミング
    - アルゴリズム
  - データの形式と変換
    - データ加工変換
    - パース
  - ツール群の分類と見つけ方
    - 配列研究関連ツールの紹介

バイオインフォマティクス独学のコツ



- 「必要」から始める
- 講座やコースを利用する
- 教則本を活用する

「必要」から始める

- 実験データのまとめ
  - 表計算ソフト＋目視・手作業
  - 追いつかない->改善の「必要」
- ツール、DB、開発環境、OS等は「とっつきやすさ」
  - 文字：ヘルプ、マニュアル、教則本
  - 人間関係：コミュニティー、周りの人

•講座やコースを利用

- 講座資料の利用
- ビデオやストーミング配信を聴講
  - JST BIRD人材養成
  - 統合データベース:統合TV
  - ライフサイエンス統合データベースプロジェクトの人材育成活動(お茶の水女子大学担当部分)
  - 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年度の演習ノート
  - 理論分子生物学(京都大学理学部)講義資料

JST BIRD
人材養成

http://lifesciencedb.jp/

ライフサイエンス統合データベ... | MotDB - MotDB

lifesciencedb.jp

LSDB 文部科学省委託研究開発事業
統合データベースプロジェクト

Guest Account | アカウント | ログイン | English

ホーム | データベース | 検索 | ツール | ダウンロード | About us

統合ホームページへようこそ

横断検索

はじめての方へ：サイトの内容をムービーやリーフレットでご紹介しています。

「統合データベースプロジェクト」とは

ポータル
- 生命科学系 データベース カタログ
- 生命科学系 学協会カタログ
- 生命科学系主要プロジェクト一覧
- 生物アイコン
- ライフサイエンス 新着論文レビュー new
- WingPro (JSTのDBポータル)
- Webリソースポータルサイト (JST解析ツールポータル)

検索
- 生命科学データベース横断検索
- 蛋白質核酸酵素 全文検索
- 文科省「ゲノム」研究報告書 全文検索
- TogoProt (蛋白質関連データベース統合検索)
- OReFiL (オンラインリソースファインダー)
- Allie (略語の正式名称を検索)
- inMeXes (文献中の英語表現を軽快に検索)

データベース
- DNAデータベース総覧と検索 (DDBJ/EMBL/GenBank)
- 遺伝子発現バンク(GEO)目次
- KazusaAnnotation & Navigation (かずさDNA研究所)
- KazusaMart (かずさDNA研究所)
- ゲノムネット医薬品データベース (京大)

アーカイブ
- 生命科学系データベースアーカイブ
- DDBJトレースアーカイブ (遺伝研 DDBJ)
- DDBJリードアーカイブ (遺伝研 DDBJ)

ツール＆解析サービス
- BodyParts3D/Anatomography
- Wired-Marker
- MiGAP (微生物ゲノムアノテーションパイプライン)
- DBCLS Galaxy

基盤技術開発
- TogoDB (誰でもデータベースが構築できる)
- TogoWS (ウェブサービスの標準化)
- OpenID 認証システム
- 統合DB情報基盤サイト (CBRC)
- 辞書の構築と公開
- LSDB Lab.
- BioHackathon(DBCLS バイオハッカソン) 2010, 2009, 2008

教材・人材育成
- 統合TV (DBやツールの動画教材)
- MotDB (教育・人材育成のサイト)

統合DB事業

新着情報
- 米澤明憲センター長が着任しました 2010-10-01 (Fri)
- 「MAGEST」(徳島大学 真壁和裕教授ら)を「生命科学系データベースアーカイブ」に追加しました 2010-9-27 (Mon)
- Medaka EST Database(東京大学 武田洋幸教授)を「生命科学系データベースアーカイブ」に追加しました 2010-9-17 (Fri)

LSDBブログ
- 施設停電に伴うサービス停止のお知らせ 2010-09-23 (Thu) 20:44:20

ニュース
- コレステロールのインプット アウトプット/制御
- 心臓発生過程で重要なIP3レセプターの役割 慶応大、理研グループ解明 先天性心疾患 理解・予防に前進
- マウスの体内にラットの膵臓 多能性幹細胞用いて作製 東大、JSTグループ成功

バナーリンク


情報とシステム2010シンポジウム

統合TV

OReFiL

http://togotv-curated.dbcls.jp/

統合TV Curated: 生命科学使い倒し系チャンネル 統合TVまとめサイト

はじめての方へ　番組コンテンツ　YET ANOTHER 統合TV　よくある質問　スタッフ　お問い合わせ

クリックして検索！

番組一覧から
キーワード検索

よく使われるツール

マウスを乗せるとツールの
簡単な説明が出ます！

- PubMed
- Entrez
- Primer3
- NCBI BLAST
- NCBI Gene Expression Omnibus
- Ensembl
- UCSC Genome Browser
- ClustalW
- BioMart
- 統計解析ソフト R
- Chimera
- アナトモグラフィー
- Jabion

使いたい・調べたい・使い倒したい
DB・ウェブツールがすぐ見つかる！
統合TVの番組探しはこちらから

ゲノム・核酸
配列解析

タンパク質
配列解析

発現制御解析

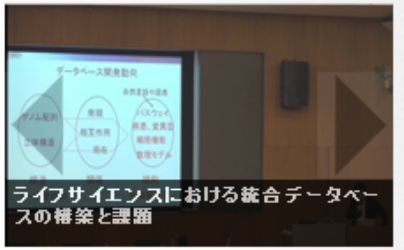
文献検索・辞書
情報収集

DBCLSサービス
講演・講習動画

データベース別
分類

Share the knowledge!

統合TVイチオシ！



最新の10番組

togotv
togotv

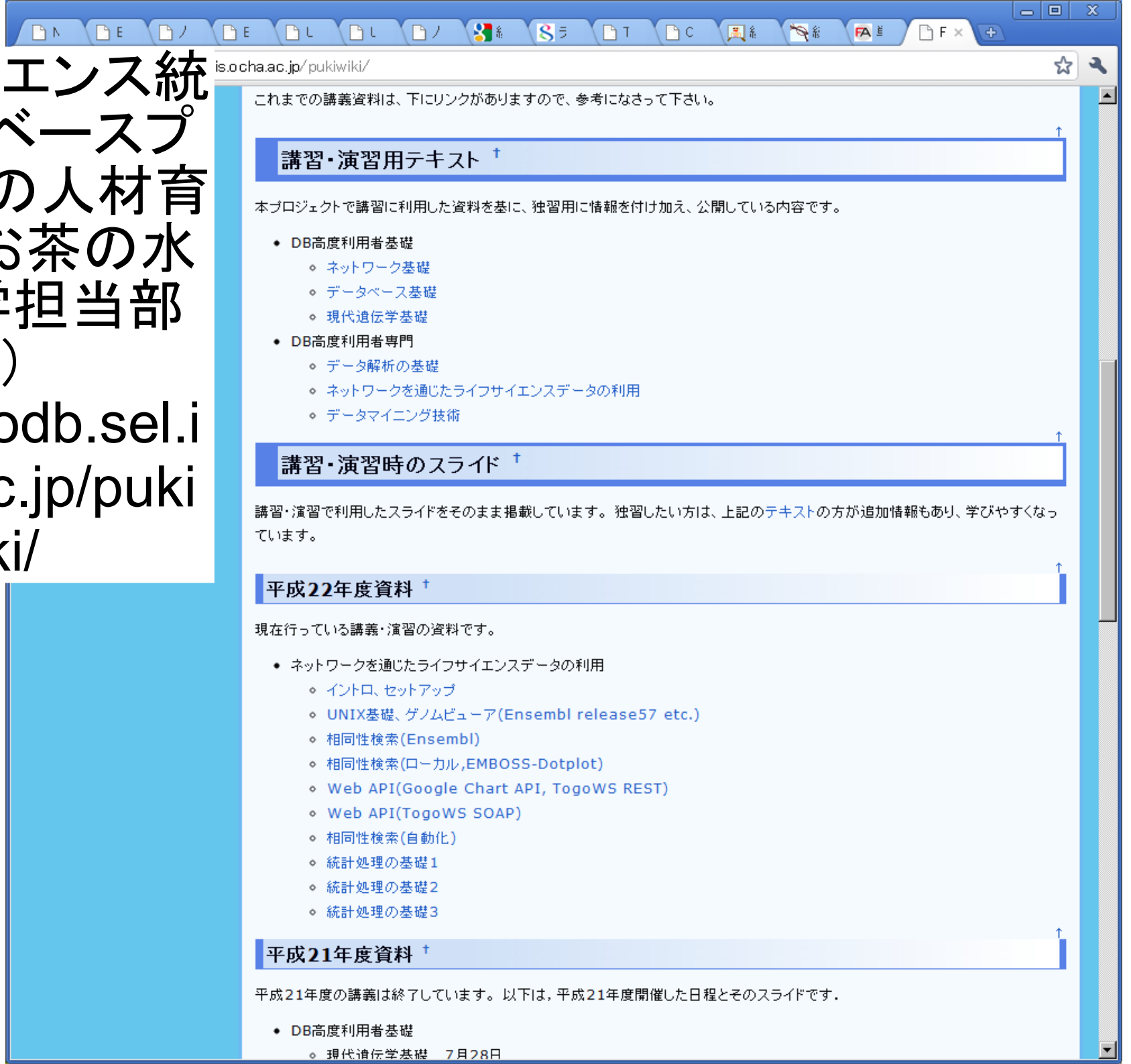
togotv [blog update] 統合TV (togotv) : [遺伝子][発現情報][可視化] GSEA softwareの使い方 ～発展編～
http://togotv.dbcls.jp/2...
35 days ago

togotv [blog update] 統合TV (togotv) : [DBCLS][IE8][English] How to use Allie 2010
http://togotv.dbcls.jp/2...
31 days ago

togotv [blog update] 統合TV (togotv) : [ゲノム][塩基配列][遺伝子] SAKURAを用いた塩基配列登録の方法(基本編)
http://togotv.dbcls.jp/2...

twitter　Join the conversation

ライフサイエンス統合データベースプロジェクトの人材育成活動（お茶の水女子大学担当部分）
http://togodb.sel.is.ocha.ac.jp/pukiwiki/



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

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

2007年度の演習ノート一覧

各回ごとに受講者の中から担当者を決め、その回のまとめノートを作ってもらうことにします。来年に新たにこの課題にチャレンジする人がこのノートを見ればほとんど内容が分かるような、そんなノートを期待しています。

- 4/6 イントロダクション
- 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CGI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

2007年度の演習ノート一覧 <--お勧め

理論分子生物学(京都大学理学部)講義資料

http://www.genome.jp/Japanese/lect/course.html

バイオインフォマティクス入門コース

理論分子生物学(京都大学理学部)講義資料

1. ゲノム解析とポストゲノム解析
   (参考資料)ヒューマンゲノム計画とニューバイオフィジックス
2. データベース技術とインターネット
   (参考資料)ゲノムネットのデータベースサービス概要
3. 分子生物学データベース
   (課題1)データベース検索(DBGET/LinkDB, PubMed)
   (課題2)タンパク質立体構造グラフィックス(RasMol, Chime)
4. 配列アライメントとホモロジー検索
   (課題3)ホモロジー検索(FASTA, BLAST)
   (課題4)マルチプルアライメント(CLUSTALW)
   (資料)式一覧
5. 構造予測と機能予測
   (課題5)モチーフ検索(MOTIF, PFAM)
   (課題6)タンパク質二次構造予測, 膜貫通部位予測(nnpredict, SOSUI)
   (課題7)タンパク質立体構造分類(SCOP, CATH)
6. ネットワーク比較とパス計算
   (課題8)パスウェイ解析(KEGG)
   (課題9)ゲノム比較解析(KEGG)
   (課題10)反応経路計算(KEGG)

問題集

- 講義レポート問題(2000～1993)
- 大学院入試問題(PDF)
  2001 | 2000 | 1999 | 1998 | 1997 | 1996 | 1995 | 1994 | 1993 | 1992 | 1991 | 1990 |

Last updated: March 24, 2001
Minoru Kanehisa

アウトライン

- [第1回]
  - 概論
    - ねらい
    - バイオインフォマティクス：どこまで把握すべきか？
  - データベース
- [第2回]
  - ゲノム関連大規模プロジェクト
  - バイオインフォマティクス独学のコツ
  - 生命科学者（実験研究者）にとってのプログラミング
    - アルゴリズム
  - データの形式と変換
    - データ加工変換
    - パース
  - ツール群の分類と見つけ方
    - 配列研究関連ツールの紹介

生命科学者にとってのプログラミング

- アルゴリズム
- プログラミングのコツと注意
- 解析作業の自動化

algorithm アルゴリズム

- (問い)アルゴリズムとは
- (例)3，7，1，9，4を昇順に並び変えよ。
- とくに考えることもなく、やればできる。(人間)
- コンピュータにやらせるには？

algorithm:並べ替え1例

- (例)3，7，1，9，4を昇順に並び変えよ。
- [方法1：基本方針]

  1. 左から順にn番目、右端をm番目とする
  2. n番目とn＋1番目を比較
  3. 右隣のほうが大きければ互いに位置を交換
  4. n＝1からn＝m－1まで繰り返す。**(<---最大値が右端)**
  5. もう一度、n＝1からn＝m－2まで繰り返す。
  6. 上記操作(5)をm－2回繰り返す。

algorithm アルゴリズム

- (問い)アルゴリズムとは
- (答え)コンピュータが実行可能な手順のこと
- 曖昧性がない
- 終了の仕方が必ず明記

生命科学者にとってのアルゴリズム

- 情報系の人たちとの打ち合わせに際し、知っていることが必要な概念
- プログラムが不得手でも、自分の考えた手順をアルゴリズムにできれば、
  - 助けがあればオリジナルなツールができる。
    - 自動化、効率化、人為ミスの解決

プログラミング言語

- プログラミング言語でアルゴリズムを記述
  - ---›コンピュータに指示をだす。
- 生命科学者にとって、プログラミング言語を学ぶべきか？
  - あなたの研究に自動化が必要ですか？
  - それはどの計算ですか、処理(判断)ですか？
  - 代表的プログラミング言語の概要だけの知識は？

生命科学者がプログラミング言語を独学するとき

- 独学の開始時(一般的に)
  - 若い方がよい
    - 遅すぎると言うことはない
  - 時代やテーマ、用途によっては、将来別の言語に乗り換えるかも
    - プログラミング言語：やること同じだからそれほど心配するな
- 情報学者やプログラマーとしての訓練を受けたものと独学者の違いは
  - あらゆるデータが渡されることを常に考える。

プログラム言語の分類

- コンパイラ方式
  - 機械語(バイナリ)に翻訳(コンパイル)してから実行
- インタプリタ方式
  - 実行時に逐次機械語に翻訳
- どちらにも分類できない言語もある

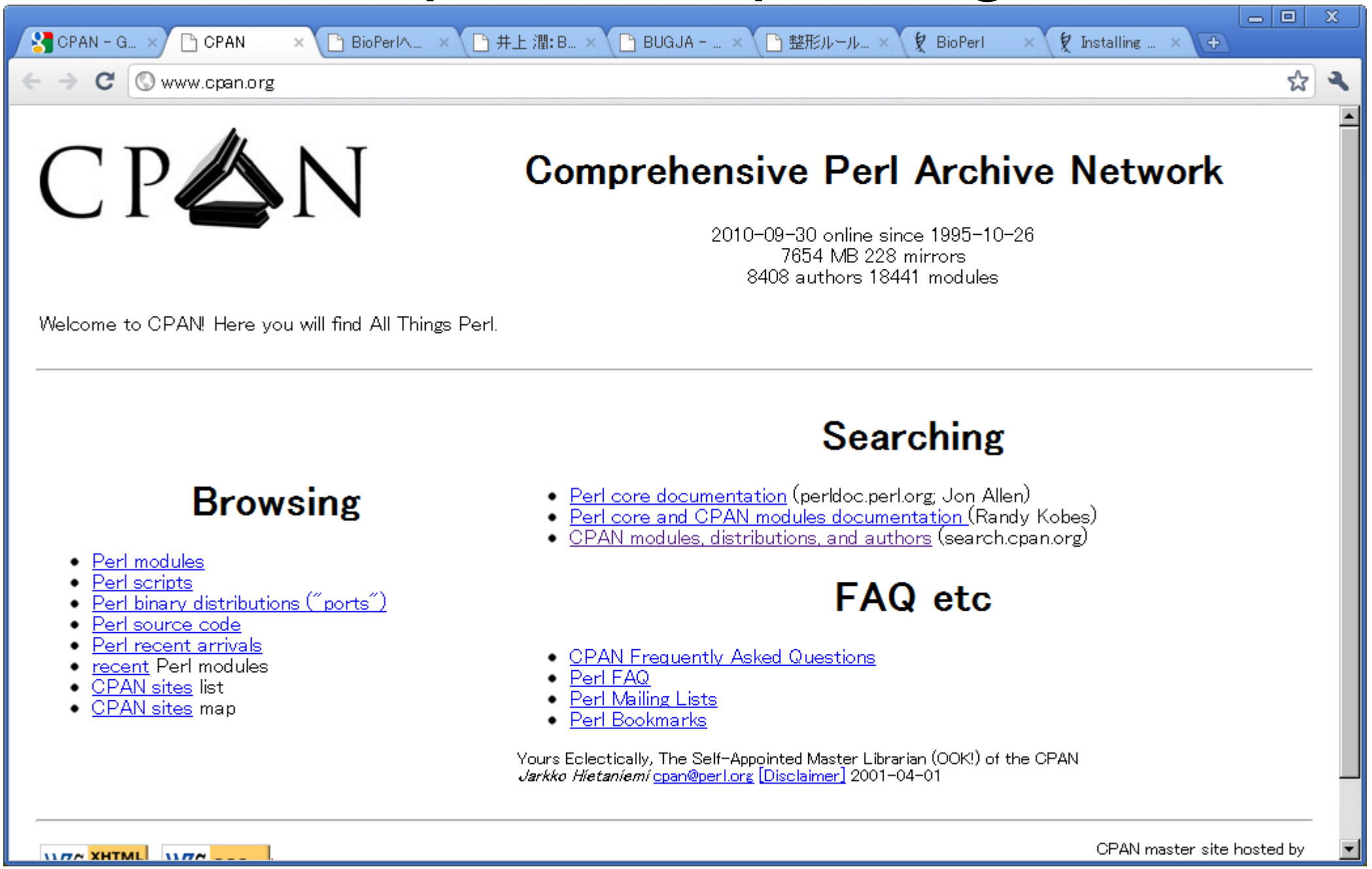
生命科学に関連深いプログラム言語

- **C/C++**
- **Java**: http://java.sun.com/
- **.NET**: http://www.microsoft.com/NET/
  - **C#**
  - **Visural Basic .NET**
- **Perl**: http://www.perl.com/
- **PHP**: http://www.php.net/
- **Python**: http://www.python.org/
- **Rubby**: http://www.ruby-lang.org/ja/
- **R言語**: http://www.r-project.org/

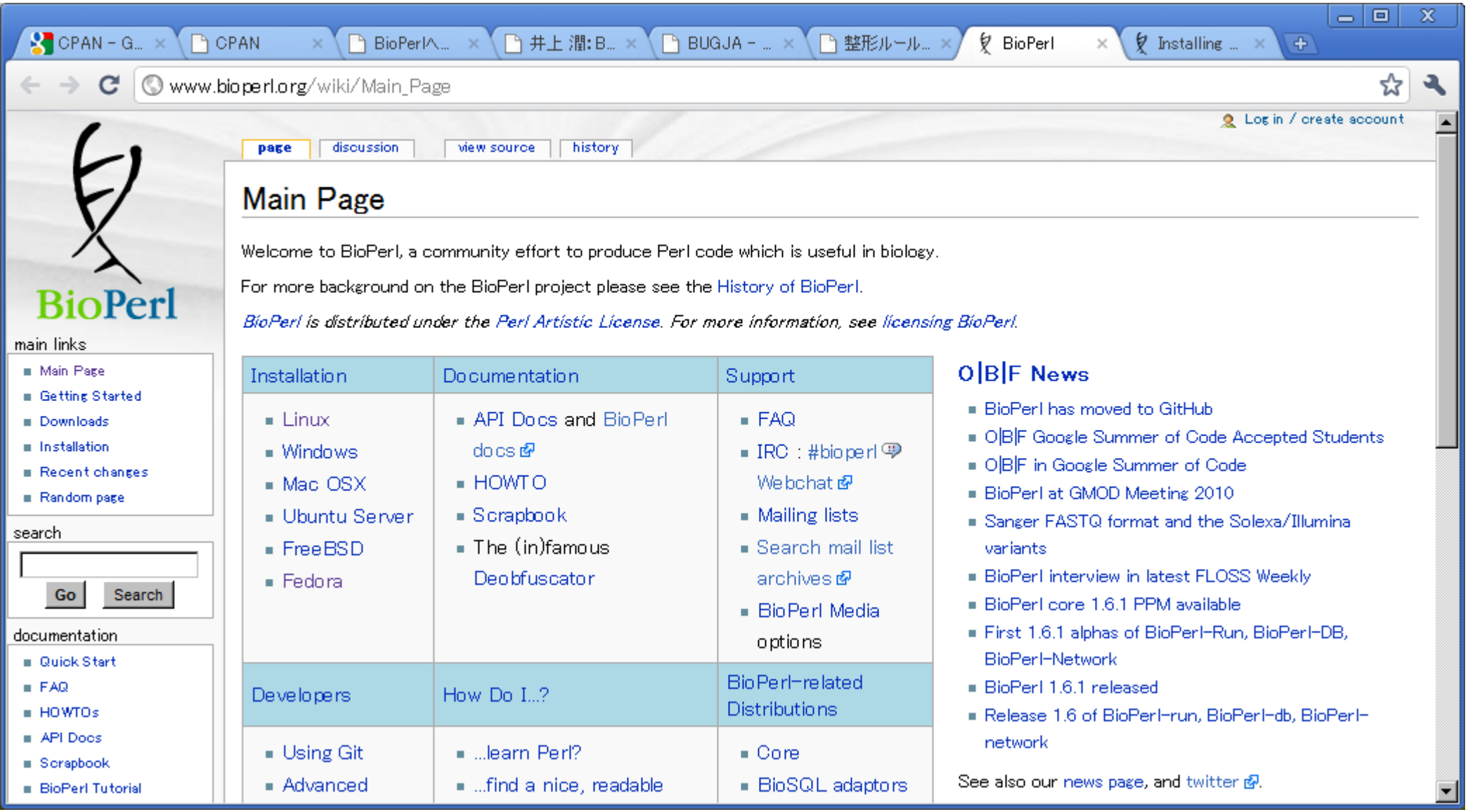
libruary (ライブラリ)とは

- 汎用性の高い複数のプログラムを、他のプログラムから利用できるように、一つにまとめたもの。
  - 関数やサブルーチンの集合。
  - ライブラリーそのものは単独で機能しない。他のプログラムの部品となる。
  - 目的と出来の良さを判断する力量が必要。
- (例)Perl言語の場合:
  - Perl言語一般のライブラリ-->CPAN
    - http://www.cpan.org/
  - 生物情報のライブラリ-->BioPerl
    - www.bioperl.org/

http://www.cpan.org/



www.bioperl.org/



生物情報のライブラリ
その他の言語

- BioPHP
  - http://www.biophp.org/
- BioRuby
  - http://bioruby.org/
- BioJava
  - http://www.biojava.org/
- BioPython
  - http://biopython.org/

解析作業の自動化

- 自動化が効果的な作業とは？
  - 長大な繰り返し作業
  - 入力、ファイルの選択、等人為的なミスが重大な作業
- 代表的な自動化方法の紹介
  - プログラム言語を利用
  - OSのコマンド処理を利用: shell script
  - マクロを利用
  - web service (ウェブ・サービス)を利用

shell script
（シェル・スクリプト）

- Linux（Mac OS Xもふくむ）
- 実行するコマンドをあらかじめテキストファイルに保存して、それを実行できる形にするもの。
- 利用者が多く、ネット上に多数の教則サイトがある。

マクロを使って自動化

- ワードやエクセルの一ファイルの中にとどまらず、アプリケーションを超えて利用できる。
- マクロ作りを比較的容易にするWindowsの支援ソフトもある。
- マクロ関連技術の発展が早く、OSのアップグレード等を契機に使いづらくなる傾向あり。

Webサービス（ウェブサービス）

- Webツール　vs. Webサービス
- Webツールは人が手で入力し、眼で理解する
- Webサービスはプログラムが入力し、理解する
- 生物情報で良く利用されるWebサービス
  API(application programming interface, アプリ開発に汎用的機能を共通利用するための手法)
  - SOAP
  - REST
  - これらは一般的にも良く利用されている
    - amazon, 楽天

生命科学系Webサービス

- EBI
  - http://www.ebi.ac.uk/Tools/webservices/
  - 英語ではあるが、丁寧なマニュアルが充実
- JST BIRD & DDBJ Web API for Biology (WABI)
  - http://xml.nig.ac.jp/index_jp.html
  - 日本語と英語のサイトがある。
- H-InvDB
  - http://hinv.jp/hinv/hws/doc/index_jp.html
- VarySysDB
  - http://hinv.jp/hinv/hws/varysysdb/doc/index_ja.html

アウトライン

- [第1回]
  - 概論
    - ねらい
    - バイオインフォマティクス：どこまで把握すべきか？
  - データベース
- [第2回]
  - ゲノム関連大規模プロジェクト
  - バイオインフォマティクス独学のコツ
  - 生命科学者（実験研究者）にとってのプログラミング
    - アルゴリズム
  - データの形式と変換
    - データ加工変換
    - パース
  - ツール群の分類と見つけ方
    - 配列研究関連ツールの紹介

データの形式と変換

- テキストデータとバイナリデータ
- 改行コード・文字コード
- データ形式：区切り型とタグ挿入型
- ゲノム情報学で良く使うデータ・フォーマット
- パース(parse)、パーサ(parser)
- 正規表現
- テキスト・エディターからプログラミングまで
- ID変換
- ゲノム・バージョン変換

テキストデータとバイナリデータ

- [問い]違いは？
- [回答例]
  - コンピュータ上のデータは全て２進数で表現できる。
  - そういった意味では全てバイナリデータ。
  - その中でもテキストとして読めるものがテキストデータ。
  - テキストとして読めるようにするのに、いろいろからくりがある。

テキストファイル：
意外と知らない改行コード

- 「CR」(Carriage Return：行頭復帰)
- 「LF」(Line Feed：改行)

タイプライタ-->テレックス：重ね打ちができるように2つを分けたのが起源

- LF:　UNIX系のシステム。Linux、Mac OS Xなど。
- CR+LF: OS/2、Microsoft Windowsなど
- CR:　Mac OS（バージョン9まで）

テキストファイル：意外と知らない文字コード（日本語）

- Shift-JIS:WindowsやMacOS（9まで）の内部コードとして使用されてきた
  - 細かくはWindowsもXp以前とVista以降では細かい部分で異なる。
- EUC-JP：Unix系の内部コードとして使用されてきた。
- UTF-7 / UTF-8：最近使われだした国際統一を目指した規格でそれぞれ7ビットと8ビット伝送路。Linuxの1distributionであるFedoraはUTF-8を使用。

おもな文字コード

文字コード名	意味
ANSI	英数字記号文字を含む、7bitの基本的な文字コード。ほかの文字コードでも、英数字部分の文字コードはこのANSI文字コードをベースにしているものがほとんどである。ASCIIコードとも呼ばれる
シフトJIS	MS-DOSの時代から広く使われている文字コード。漢字文字コードとして、「JISコード」を変形させたもの（シフトしたもの）を利用して、ANSI文字と共存させているのでこう呼ばれる。PC環境では一般的な日本語文字コード
EUC	UNIX環境で広く使われている日本語文字コード。シフトJISとは異なる方法でANSI文字と漢字文字を共存させている
JISコード	ANSI文字コードと漢字文字コードを、「エスケープ・シーケンス」と呼ばれる特別な文字シーケンスで切り替えながら共存させている。ほとんどの場合、インターネット電子メールやニュースは、この文字コードで送受信される
Unicode	世界中の文字を16bitもしくは32bitの固定長の文字コードで統合的に扱うために作られた文字コード。文字の種類によらずコード長が一定しているので、プログラムから扱いやすく、OSやアプリケーションの内部コードとして使われることが多い。ただし次のUTF-16と同義に使われる場合も多い
UTF	Unicodeをベースにして、実際にファイルに格納したり、通信を行う場合のバイト・データの並べ方などを規定したもの。UTF-8とかUTF-16などがある。UTF-16では、バイト・オーダーの違いにより、リトル・エンディアン形式とビッグ・エンディアン形式などの違いがある

http://www.atmarkit.co.jp/fwin2k/win2ktips/395codeconv/codeconv.html

テキストファイル：意外と知らない文字コード（まとめ）

- 時代の流れ
  - 初期：American National Standards Institute(ANSI)　英語のアルファベットと記号を表現するもの。
  - 一昔前：各国語でそれぞれに対応していた時代。
  - 最近：統一規格で各国語に対応できるコードを。
- 実際使用に際して
  - 汎用機<--->パソコン、OS間でファイルをやり取りする
    - 自動で変換されない場合は、要変換
      - 開くソフトでエンコードを変更。
      - または専用変換ソフトを使う。
    - ファイル名は普段から2バイト文字を使わないように**注意**。

データ形式:タグ挿入型と区切り型

XML形式	HTML形式	CSV形式	フラットファイル形式
<成績表> <名前>鈴沢 一郎</名前> <国語>67</国語> <算数>70</算数> <理科>95</理科> <社会>87</社会> </成績表>	<TABLE> <TD>鈴沢 一郎</TD> <TD>67</TD> <TD>70</TD> <TD>95</TD> <TD>87</TD> </TABLE>	鈴沢 一郎, 67, 70, 95, 87	名前　鈴沢 一郎 国語　67 算数　70 理科　95 社会　87

- Markup Language(ML)では項目をタグで囲ってあらわす。
- (ML)階層的なデータ構造定義が可能
- XML（eXtensible Markup Language）はタグを自由に設定できる。
- (ML)Document Type Definition(DTD)で要素の型と、要素の親子関係を定義
- XMLはXSLスタイルシートにより、データの内容と表現を分離して管理可能
- 区切り型はComma Separated Values (CSV), Tab SV(TSV)

Style sheet

- 元々の意味:
  - フォントの種類や文字の大きさ、色、行間の幅、修飾など、文書の見栄えに関する情報をひとまとめにした、文書の雛形のこと
- HTML文書にスタイルシートを適用する
  - CSSという言語を使用
  - 複数のページの見栄えを統一することができる。
- HTML文書からレイアウト情報などを分離。
- HTML文書の論理的な構造が把握しやすくなる。
- XML文書の場合は、XSLという言語を使用。

ゲノム情報学で良く使われる データ・フォーマット

- 配列用フォーマット
  - 単独
  - アライメント
  - 例）FASTA,PHYLIP,ALN
- ゲノム領域用フォーマット
  - 例）GFF

•データ入出力はツールに依存する。まずは従うことが肝要。
•互いに変換可能。
  • 文字列扱いに便利なプログラム言語が多様されるようになった。

代表的な配列ファイルフォーマット
FASTA (format, file)

```
>seq_1
CTCCATAATCAT
>seq_2
CTCCATAATTCAT
```

```
>seq_1
CTCCATAAT-CAT
>seq_2
CTCCATAATTCAT
```

代表的な配列ファイルフォーマット
PHYLIP (format, file)

```
 2 357
p0Sj_1          CTCCATAAT- CATCACTATA CACAATTAAC GAATTCTAAA TAC------A
p0Sp_1          CTACATAATA CTTCAATAAC CTAAATGATC GAAATTTCAA TAATTTTTAA

                TCACTATTGT ATCCACTGTA TGCACGTCAC AAAAGTGCGT CGCTGTGACT
                TGACTATTTG ACATTCTTTA ---ACAGTAT ACTAGGGCAG GGTTGTGACT

                GTCTTTTTCT TCCTATATAA GCAA---CTC GAAAC----- -------GTC
                GTCCAAGGAT T---ATATAA GTAGAAGCTC ATAACAAGAG GTGGTTTGTC

                TCCTAGAGAG CGTCATCCAT CAAAGGTAAG C---AGTTCT GAATAAGGAA
                AATTACAAAA CAGAATGTAA GTAATTTCAA CTTGAGTTTT AAAACGCTGA

                CACAGCTTGT ACAGTAATC- -ATGATGAGT AACT-GTGTC GCCCCAATCC
                AACAATGTTA AAAATATTAC TATGATGAGC AATTTGTGTC GCCCCTATCC

                GAAACATAGT GATGATTGAC CTGGACGCCA AATCCACT-T CTGATCGTTA
                GAAATCT-GT GACGATTGAC CCGGAAGCCA AATCCGATCT CTGAATATTT

                CTGTACATTC A-GTGATTCT ---TAACATC CT-TTATGTG CACTCTTTAC
                CTATTGTTTC AAGTGCTTTT CCATGAAATC TTATTATTTG AATAATTTAC

                TAACTAT
                TAACTTT
```

代表的な配列ファイルフォーマット (Clustal W) ALN (format, file)

```
CLUSTAL W (1.83) multiple sequence alignment


Seq_03              CTACATAATACTGCAATAAGCTAAATGATCGAAATTTCAA
CU329672-1          CTACATAATACTTCAATAACCTAAATGATCGAAATTTCAA
                    ********** ****** ********************
```

ゲノム領域データのフォーマット

- 単一配列やalignment配列のフォーマットは配列の文字列やそれらのalignment関係
- ゲノム領域はDNAストランド、ゲノム座標等
- MAFは両者を合わせたフォーマット

BED format
bigBed format
PSL format
GFF format
GTF format
MAF format
BAM format
WIG format
bigWig format
Microarray format
Chain format
Net format
Axt format
.2bit format
.nib format
GenePred table format

GFF (v3) format

```
##gff-version   3
##sequence-region   ctg123 1 1497228
ctg123 . gene    1000  9000  .  +  .  ID=gene1;Name=EDEN

ctg123 . TF_bind 1000  1012  .  +  .  ID=tfbs1;Parent=gene1

ctg123 . mRNA    1050  9000  .  +  .  ID=mRNA1;Parent=gene1;Name=EDEN.1
ctg123 . mRNA    1050  9000  .  +  .  ID=mRNA2;Parent=gene1;Name=EDEN.2
ctg123 . mRNA    1300  9000  .  +  .  ID=mRNA3;Parent=gene1;Name=EDEN.3
ctg123 . exon    1300  1500  .  +  .  ID=exon1;Parent=mRNA3
ctg123 . exon    1050  1500  .  +  .  ID=exon2;Parent=mRNA1,mRNA2
ctg123 . exon    3000  3902  .  +  .  ID=exon3;Parent=mRNA1,mRNA3
ctg123 . exon    5000  5500  .  +  .  ID=exon4;Parent=mRNA1,mRNA2,mRNA3
ctg123 . exon    7000  9000  .  +  .  ID=exon5;Parent=mRNA1,mRNA2,mRNA3
ctg123 . CDS     1201  1500  .  +  0  ID=cds1;Parent=mRNA1;Name=edenpr.1
ctg123 . CDS     3000  3902  .  +  0  ID=cds1;Parent=mRNA1;Name=edenpr.1
```

1:seqid
2:source
3:type 4:start 5:end 6:score
7:strand
8: frame
9:attributes

parse(パース、構文解析)

- 情報学では構文解析は字句解析とプログラムの文法の正しさの判断との2つの部分を指す。
  - うらにわにはにわにわにはにわにわとりがいる
  - 裏庭には庭庭には庭鶏がいる。
  - 裏庭に葉二把、庭には丹羽鶏がいる。
  - 裏にワニ埴輪、庭には丹羽鶏がいる。
  - 裏庭に埴輪、庭に埴輪、鶏がいる。
  - 裏庭には二羽、庭には二羽、鶏がいる。

parse(パース、構文解析)

- 情報学では構文解析は字句解析とプログラムの文法の正しさの判断との2つの部分を指す。
- XMLファイルでは、parserを介して、テキスト部分を抜き出すことにより、人の目で解釈できる。
- バイオインフォマティクスでは:
  - BLAST結果などヒトが眼でみるために書かれたファイル
  - 字句解析-->次の処理(プログラムでは)
  - いくつかのライブラリとして公開されている
    - BioPerlでBLASTをparseするライブラリ
    - http://www.bioperl.org/wiki/Parsing_BLAST_HSPs

<Hsp_hit-frame>0</Hsp_hit-frame>
<Hsp_identity>103</Hsp_identity>
<Hsp_positive>170</Hsp_positive>
<Hsp_gaps>21</Hsp_gaps>
<Hsp_align-len>333</Hsp_align-len>

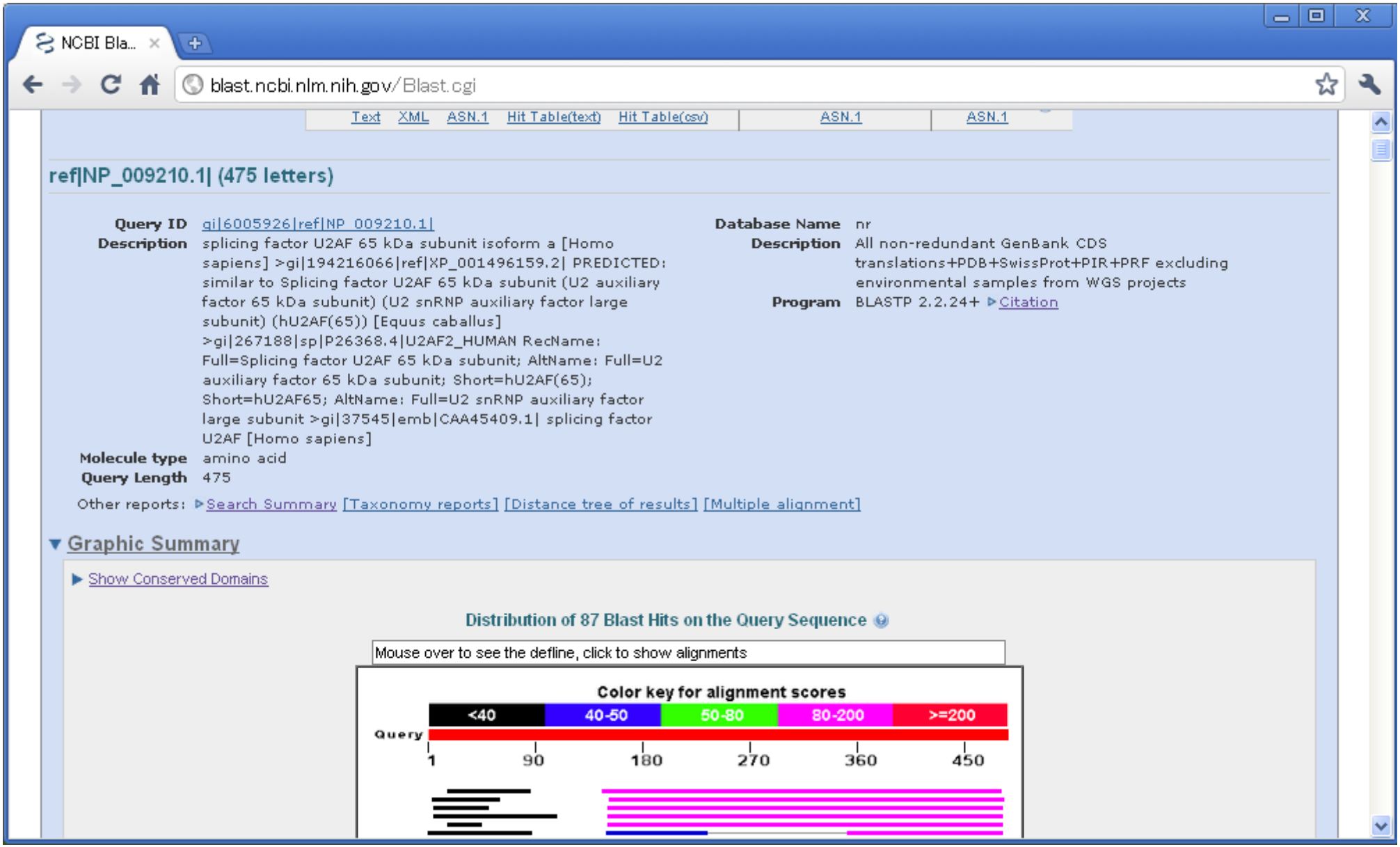
<Hsp_qseq>QARRLYVGNIPFGITEEAMMDFFNAQMRLGGLTQAPGNPVLAVQINQDKNFAFLEFR
SVDETTQAMAFDGIIFQGQSLKIRRPHDYQPLP---
GMSENPSVYVPGVVSTVVPDSAHKLFIGGLPNYLNDDQVKELLTSFGPLKAFNLVKDSATGLSKGY
AFCEYVDINVTDQAIAGLNGMQLGDKKLLVQRASVGAKNATLVSPPSTINQTPVTLQVPGLMSSQV
QMGGH--
PTEVLCLMNMVLPEELLDDEEYEEIVEDVRDECSKYGLVKSIEIPRPVDGVEVPGCGKIFVEFTSVF
DCQKAMQGLTGRKFANRVVVTKYCDPDSYHRRDF</Hsp_qseq>

<Hsp_hseq>HAVRLYLGNLPDNVDKDHLHNYIRQQMESHGAVLDPGDPVIQVQLQPGQKYCFVQF
RSIEETEAALQIDTINYQGKPLKFKRVKDYEISPRIEGEREVPKIQ--------
PKEPAQKLFVCGLAPDTDNDALANILSEYGNLKSLNVVRD-
IKNVCKGFAFCEFETDLETQNCVNGLNNKVIGGRLLQVK------
KNAQLPTPTQDYIIDTITLGEQSAFEAKLQQINQMKVSSVVVINNAVRIKNIEDDYEYNFIVKDLKKEI
EKIGRLISMVVPRKKDGYS-
EGIGKVFVEFENEQFAKIAIILLQNKKYDGREIDIAFYDPRLYADKQY</Hsp_hseq>
<Hsp_midline> A RLY+GN+P + ++ + ++ QM G PG+PV+ VQ+ + + F++FRS++ET A+ D I
+QG+ LK +R DY+ P G E P + + A KLF+ GL ++D + +L+ +G LK+ N+V+D + KG+AFCE+
T + GLN +G + L V+ KNA L +P +TL +++Q + V+ + N V + + DD EY IV+D++ E K
G + S+ +PR DG G GK+FVEF + + A+ L +K+ R + + DP Y + +</Hsp_midline>
</Hsp>
</Hit_hsps>
</Hit>
<Hit>

```
Database: All non-redundant GenBank CDS
translations+PDB+SwissProt+PIR+PRF excluding environmental samples
from WGS projects
           11,974,163 sequences; 4,087,290,020 total letters
Query= gi|6005926|ref|NP_009210.1| splicing factor U2AF 65 kDa subunit
isoform a [Homo sapiens] >gi|194216066|ref|XP_001496159.2| PREDICTED:
similar to Splicing factor U2AF 65 kDa subunit (U2 auxiliary factor 65
kDa subunit) (U2 snRNP auxiliary factor large subunit) (hU2AF(65))
[Equus caballus] >gi|267188|sp|P26368.4|U2AF2_HUMAN RecName:
Full=Splicing factor U2AF 65 kDa subunit; AltName: Full=U2 auxiliary
factor 65 kDa subunit; Short=hU2AF(65); Short=hU2AF65; AltName:
Full=U2 snRNP auxiliary factor large subunit >gi|37545|emb|CAA45409.1|
splicing factor U2AF [Homo sapiens]
Length=475


                                                                  Score     E
Sequences producing significant alignments:                      (Bits)  Value

ref|XP_001459639.1|  hypothetical protein [Paramecium tetraure...   156    3e-38
ref|XP_001457804.1|  hypothetical protein [Paramecium tetraure...   151    1e-36
ref|XP_001454069.1|  hypothetical protein [Paramecium tetraure...   149    4e-36
```


NCBI Bla...
blast.ncbi.nlm.nih.gov/Blast.cgi
Text XML ASN.1 Hit Table(text) Hit Table(csv)
ASN.1
ASN.1
ref|NP_009210.1| (475 letters)
Query ID gi|6005926|ref|NP_009210.1|
Description splicing factor U2AF 65 kDa subunit isoform a [Homo sapiens] >gi|194216066|ref|XP_001496159.2| PREDICTED: similar to Splicing factor U2AF 65 kDa subunit (U2 auxiliary factor 65 kDa subunit) (U2 snRNP auxiliary factor large subunit) (hU2AF(65)) [Equus caballus] >gi|267188|sp|P26368.4|U2AF2_HUMAN RecName: Full=Splicing factor U2AF 65 kDa subunit; AltName: Full=U2 auxiliary factor 65 kDa subunit; Short=hU2AF(65); Short=hU2AF65; AltName: Full=U2 snRNP auxiliary factor large subunit >gi|37545|emb|CAA45409.1| splicing factor U2AF [Homo sapiens]
Molecule type amino acid
Query Length 475
Database Name nr
Description All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Program BLASTP 2.2.24+ Citation
Other reports: Search Summary [Taxonomy reports] [Distance tree of results] [Multiple alignment]
Graphic Summary
Show Conserved Domains
Distribution of 87 Blast Hits on the Query Sequence
Mouse over to see the defline, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
90
180
270
360
450

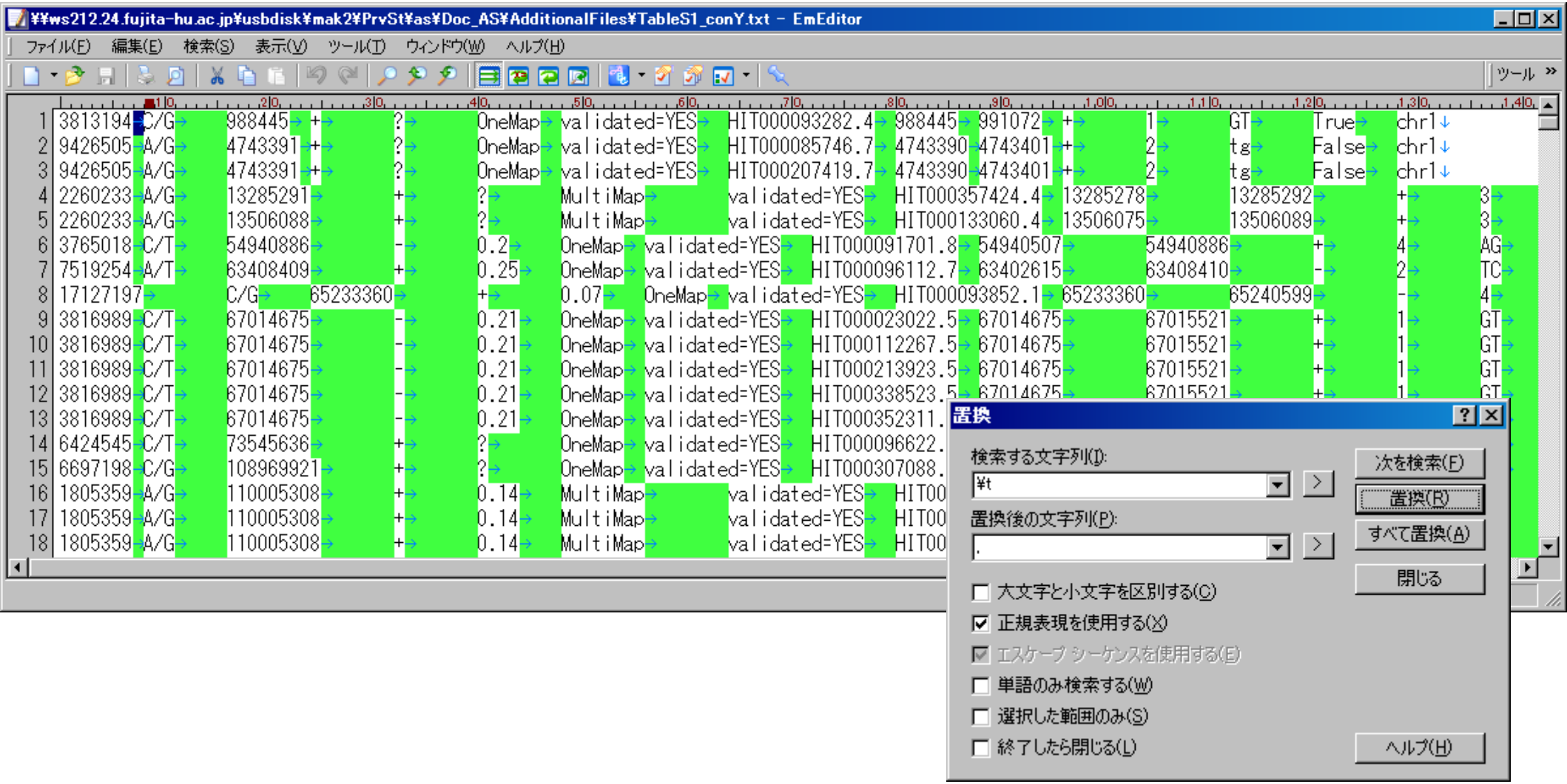
正規表現 (regular expression)

- 文字列の集合を一つの文字列で表現する方法。
- 正規表現を使うと、文字列の検索や置換をパターンで行う事ができる。
- 配列のデータは文字列のデータであるので、検索だけでなく、プログラミングにも正規表現は重要。
- 例）
  - ?：直前の文字が0か1個ある。"*colou?*r" は、
  - *：直前の文字が0個以上ある。"*go*gle*" は、
  - |：またはの意。
  - \dまたは[0-9]：数字にマッチ。
  - \n：改行コードにマッチ。

要チェックツール：テキストファイル作成

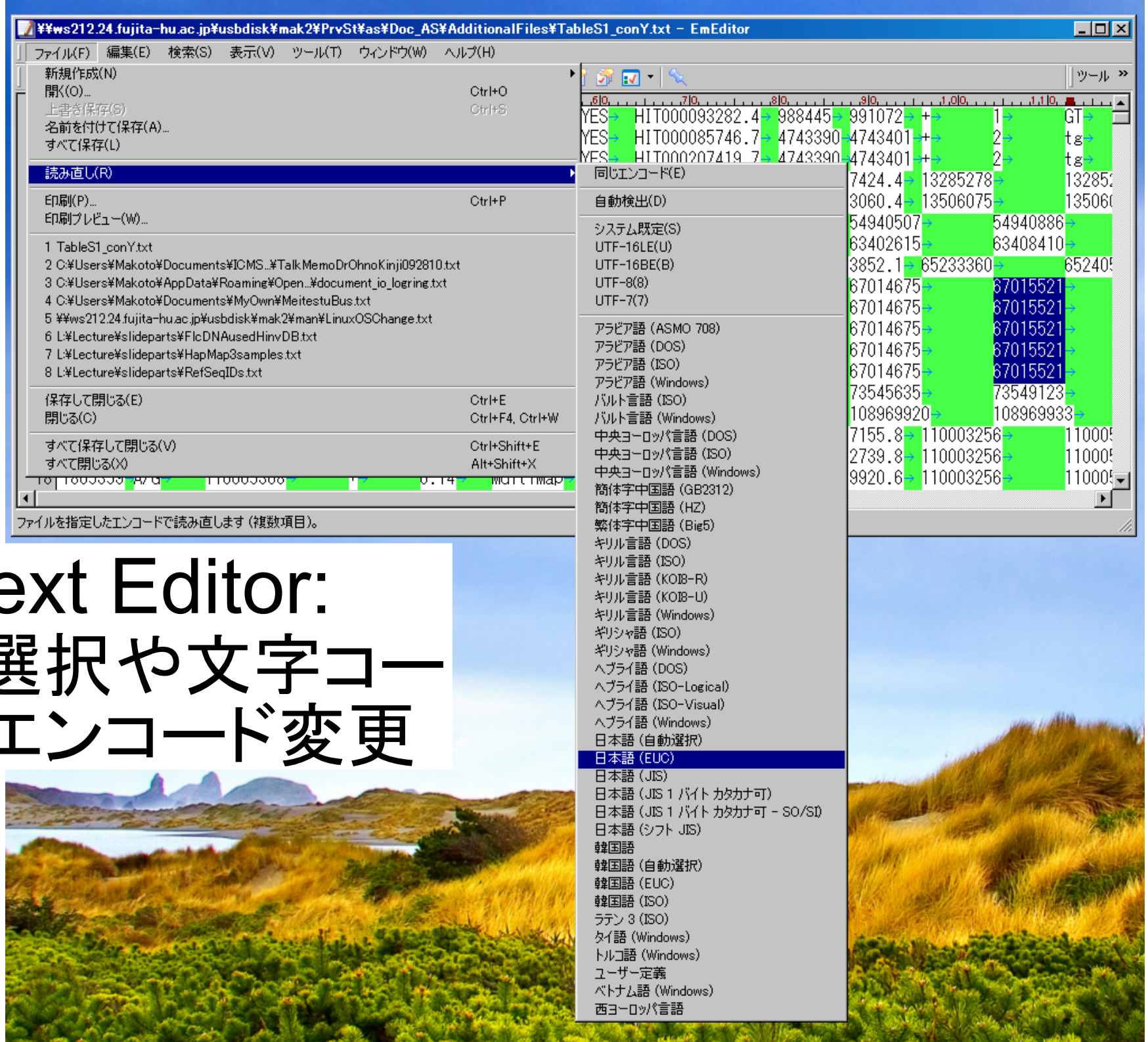
- テキスト・エディター
  - OSに付属しているものでも可能だが、正規表現使えますか？

テキスト・エディタ
正規表現を使えるか



要チェックツール：テキストファイル作成

- テキスト・エディター(便利な機能)
  - OSに付属しているものでも可能だが、正規表現使えますか？
    - タブ区切り<-->コンマ区切りなど表の整形に便利
    - 意図しない改行を一括除去
  - 縦型の選択-->コピーやカット
  - 文字コードを変更して読みなおし



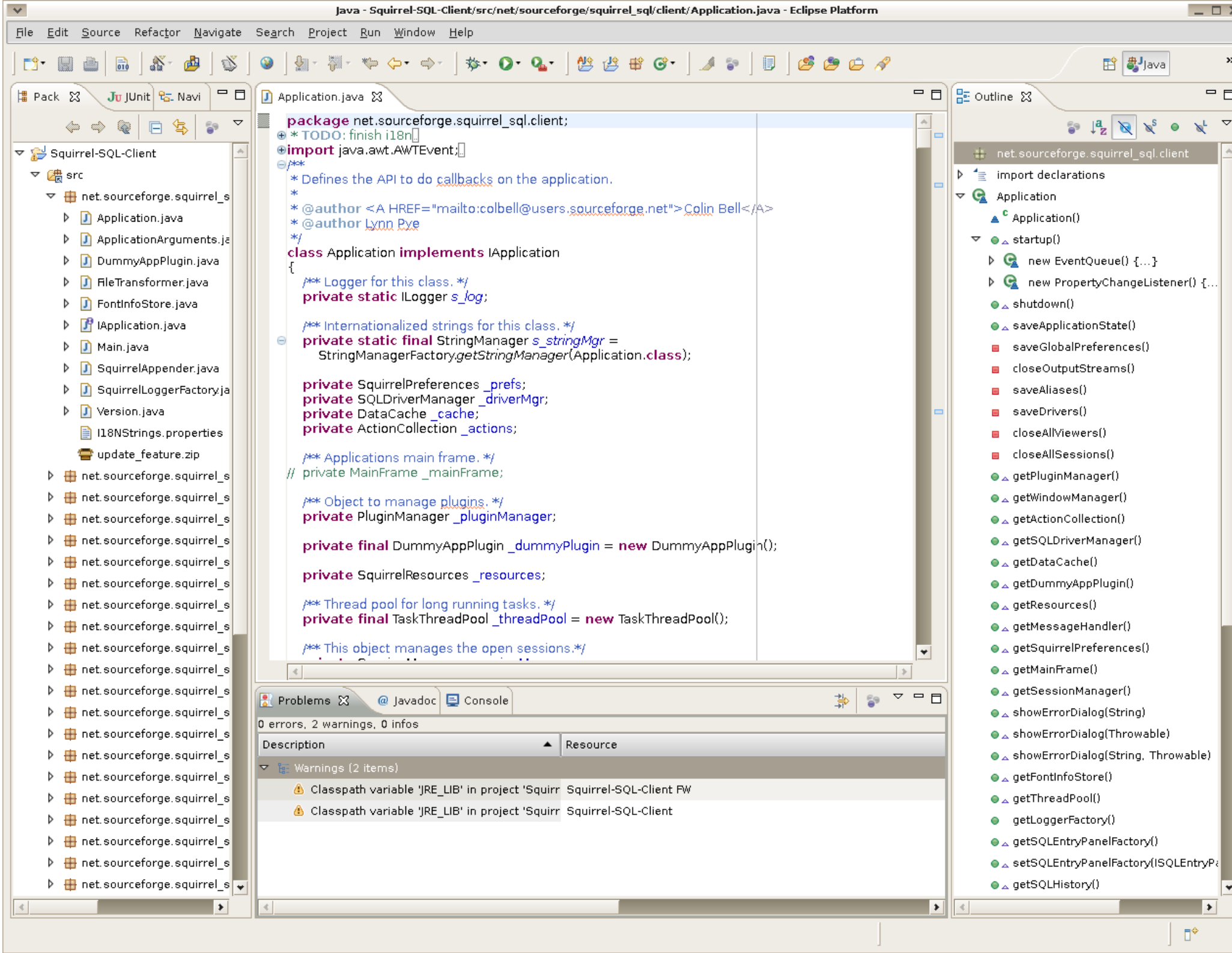
Text Editor: 縦型選択や文字コードのエンコード変更

要チェックツール：テキストファイル作成

- テキスト・エディター
  - 正規表現、縦型選択、コード変更などの機能が使える。
  - プログラミング支援機能付きはプログラム・エディターともよばれる
    - タグ挿入型：色違いで表示
  - 選ぶポイント：ショートカットキー
    - （新規習得か既知にこだわるか）
  - 例：秀丸、vi, Emacs, Gedit

Text Editorと統合開発環境(IDE)

- 統合開発環境
  - プログラム作成の際に必要な、テキスト・エディタ、コンパイラ、デバッガなどのツールを一つにしたもの。
  - 例：Eclips (IBM), KDevelop (KDE)
  - EmacsのようなText editorでもマクロを使うことでIDEとなる。
- （[参考]XML Parser: 実績あるものを選ぶ必要<---プログラマー）




```
package net.sourceforge.squirrel_sql.client;
* TODO: finish i18n
import java.awt.AWTEvent;
/**
 * Defines the API to do callbacks on the application.
 *
 * @author <A HREF="mailto:colbell@users.sourceforge.net">Colin Bell</A>
 * @author Lynn Pye
 */
class Application implements IApplication
{
   /** Logger for this class. */
   private static ILogger s_log;

   /** Internationalized strings for this class. */
   private static final StringManager s_stringMgr =
      StringManagerFactory.getStringManager(Application.class);

   private SquirrelPreferences _prefs;
   private SQLDriverManager _driverMgr;
   private DataCache _cache;
   private ActionCollection _actions;

   /** Applications main frame. */
//  private MainFrame _mainFrame;

   /** Object to manage plugins. */
   private PluginManager _pluginManager;

   private final DummyAppPlugin _dummyPlugin = new DummyAppPlugin();

   private SquirrelResources _resources;

   /** Thread pool for long running tasks. */
   private final TaskThreadPool _threadPool = new TaskThreadPool();

   /** This object manages the open sessions.*/
```

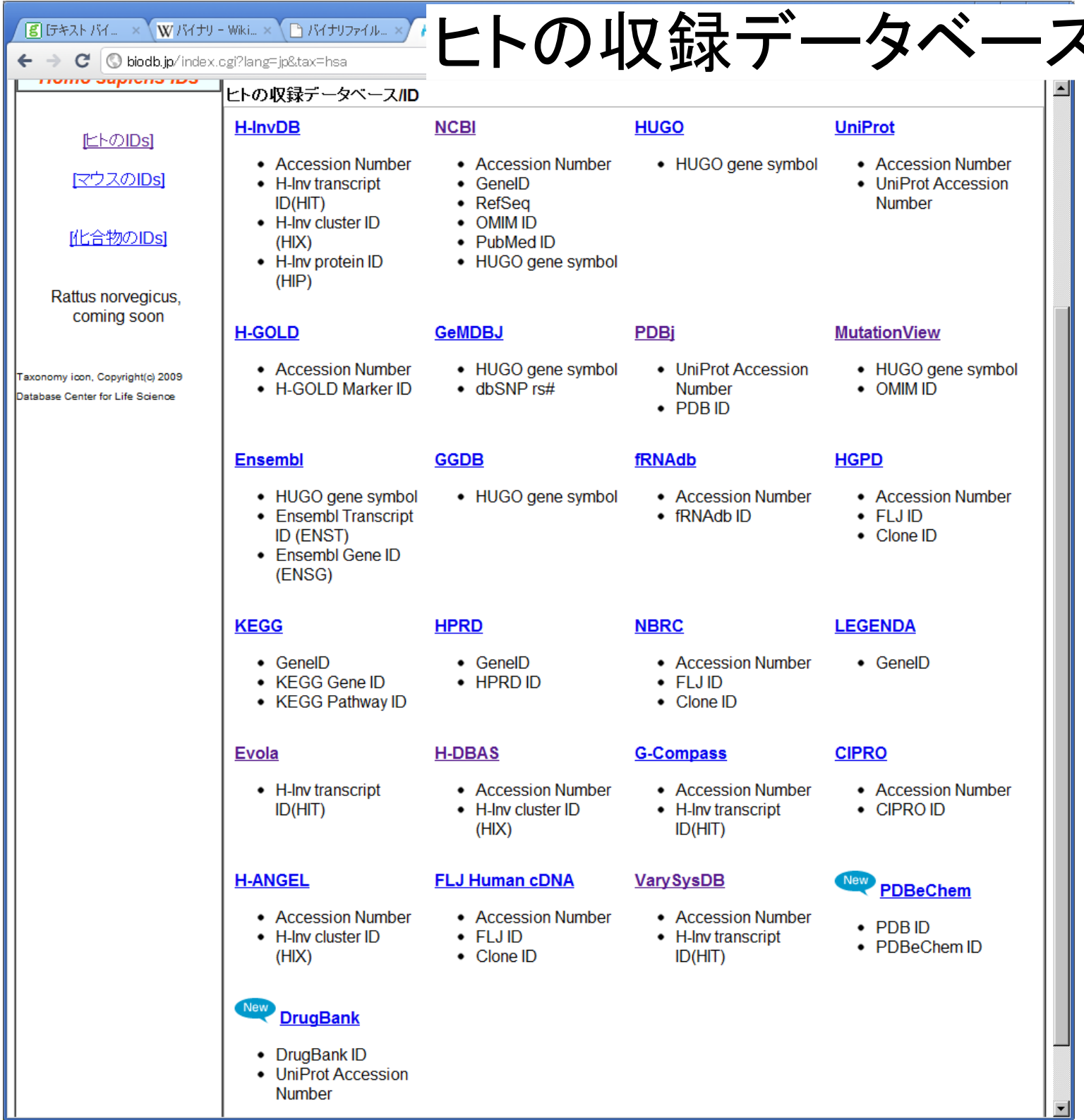
文字列の変換以外の良く使う変換

- IDの変換
- ゲノム位置の変換
  - 同種のゲノム（アッセンブリ）バージョン間
  - 異種間

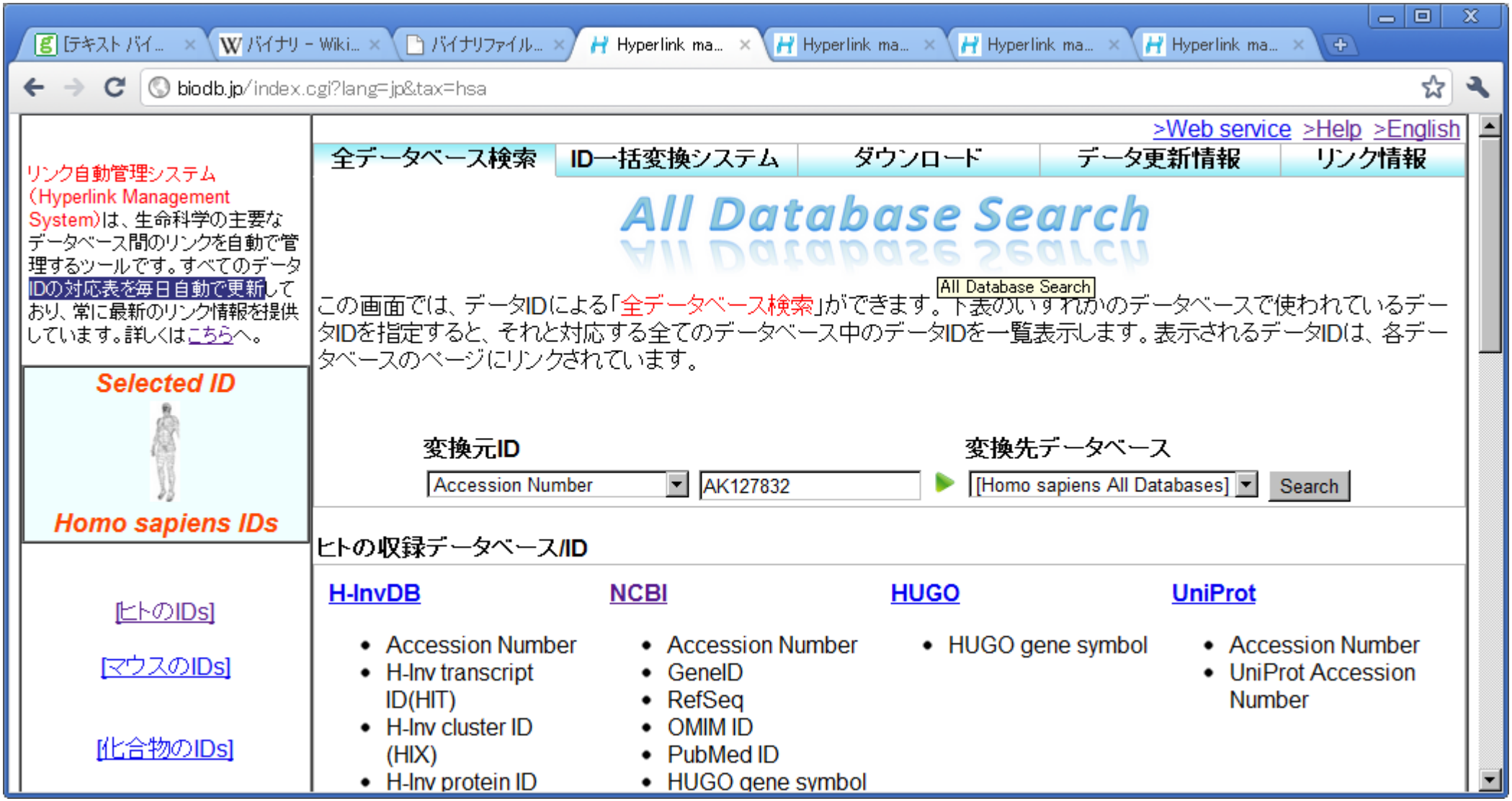
ID変換

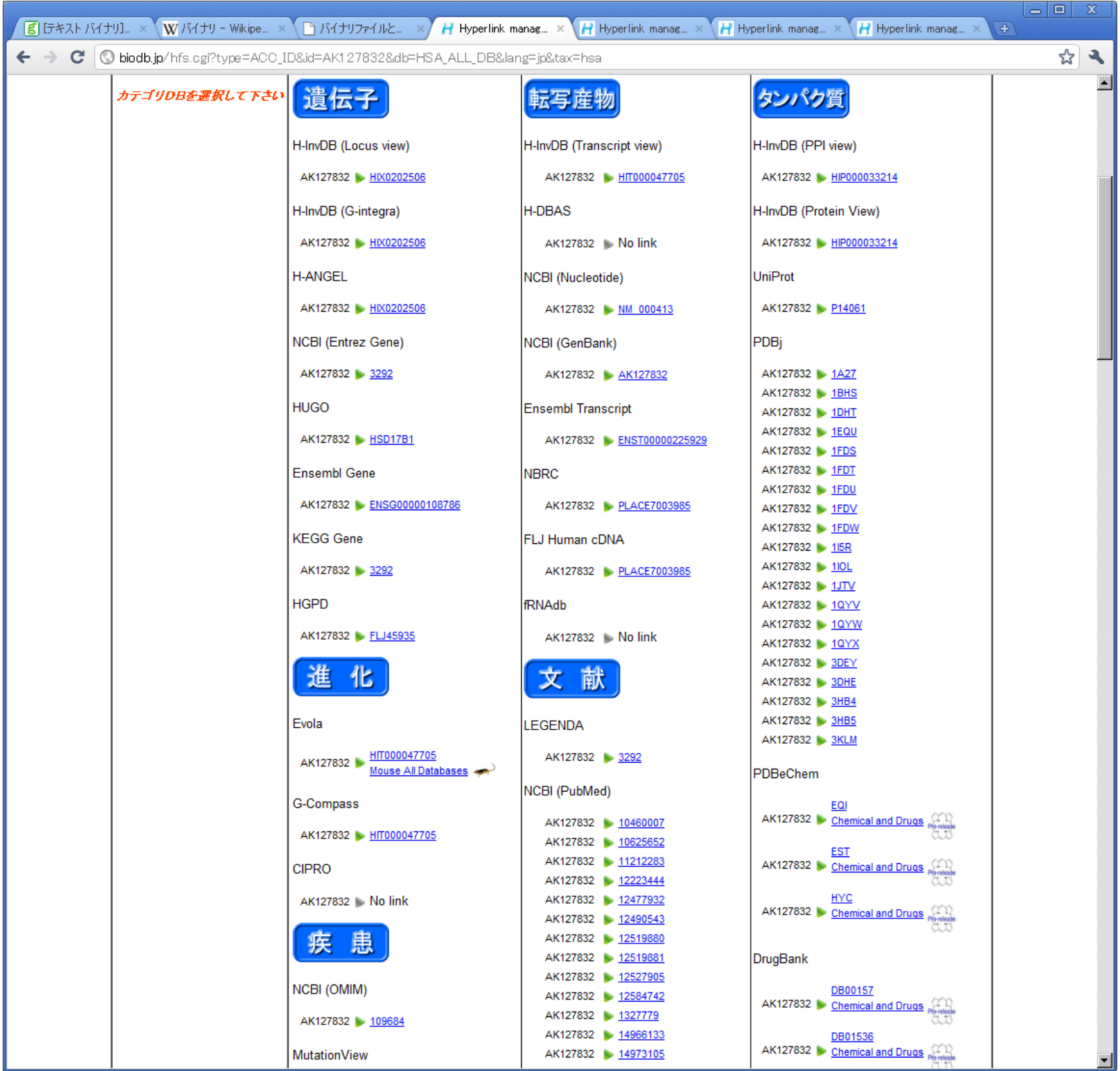
- リンク自動管理システム（Hyperlink Management System, http://biodb.jp）
  - 生命科学の主要なデータベース間のリンクを自動で管理するツール。
  - IDの対応表を毎日自動で更新。
  - webツールを使う。
  - プログラムから利用する。

ヒトの収録データベースとID

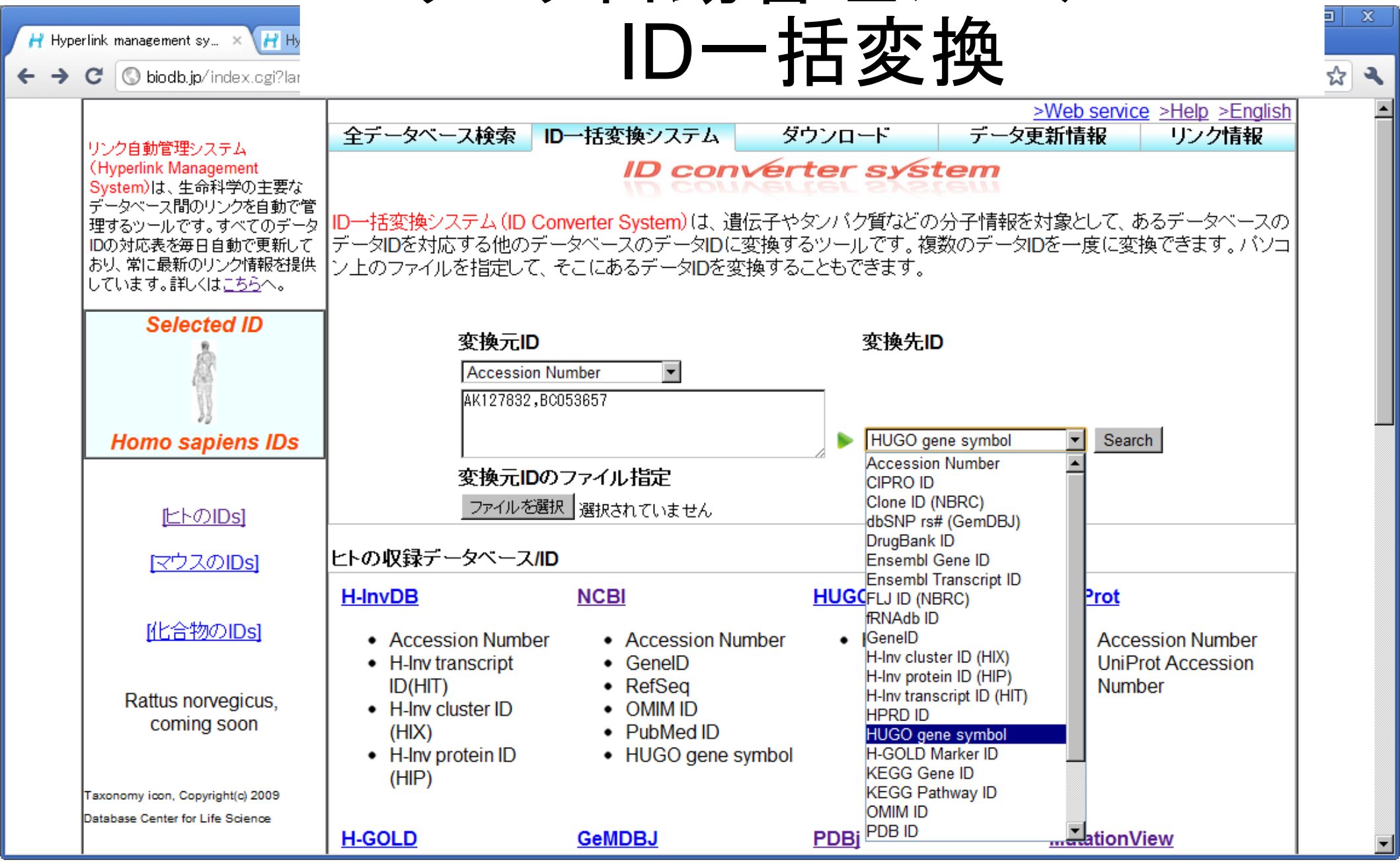


リンク自動管理システム：全データベース検索




[テキスト バイナリ]...
バイナリ - Wikipe...
バイナリファイルと...
Hyperlink manag...
Hyperlink manag...
Hyperlink manag...
Hyperlink manag...
biodb.jp/hfs.cgi?type=ACC_ID&id=AK127832&db=HSA_ALL_DB&lang=jp&tax=hsa
カテゴリDBを選択して下さい
遺伝子
H-InvDB (Locus view)
AK127832 HIX0202506
H-InvDB (G-integra)
AK127832 HIX0202506
H-ANGEL
AK127832 HIX0202506
NCBI (Entrez Gene)
AK127832 3292
HUGO
AK127832 HSD17B1
Ensembl Gene
AK127832 ENSG00000108786
KEGG Gene
AK127832 3292
HGPD
AK127832 FLJ45935
進 化
Evola
AK127832 HIT000047705
Mouse All Databases
G-Compass
AK127832 HIT000047705
CIPRO
AK127832 No link
疾 患
NCBI (OMIM)
AK127832 109684
MutationView
転写産物
H-InvDB (Transcript view)
AK127832 HIT000047705
H-DBAS
AK127832 No link
NCBI (Nucleotide)
AK127832 NM_000413
NCBI (GenBank)
AK127832 AK127832
Ensembl Transcript
AK127832 ENST00000225929
NBRC
AK127832 PLACE7003985
FLJ Human cDNA
AK127832 PLACE7003985
fRNAdb
AK127832 No link
文 献
LEGENDA
AK127832 3292
NCBI (PubMed)
AK127832 10460007
AK127832 10625652
AK127832 11212283
AK127832 12223444
AK127832 12477932
AK127832 12490543
AK127832 12519880
AK127832 12519881
AK127832 12527905
AK127832 12584742
AK127832 1327779
AK127832 14966133
AK127832 14973105
タンパク質
H-InvDB (PPI view)
AK127832 HIP000033214
H-InvDB (Protein View)
AK127832 HIP000033214
UniProt
AK127832 P14061
PDBj
AK127832 1A27
AK127832 1BHS
AK127832 1DHT
AK127832 1EQU
AK127832 1FDS
AK127832 1FDT
AK127832 1FDU
AK127832 1FDV
AK127832 1FDW
AK127832 1I5R
AK127832 1IOL
AK127832 1JTV
AK127832 1QYV
AK127832 1QYW
AK127832 1QYX
AK127832 3DEY
AK127832 3DHE
AK127832 3HB4
AK127832 3HB5
AK127832 3KLM
PDBeChem
AK127832 EQI Chemical and Drugs
AK127832 EST Chemical and Drugs
AK127832 HYC Chemical and Drugs
DrugBank
AK127832 DB00157 Chemical and Drugs
AK127832 DB01536 Chemical and Drugs

リンク自動管理システム：ID一括変換



リンク自動管理システム：プログラムから利用する

2. 自動管理によるリンクの設定方法

(1) 単純なリンクの方法

本サービスは、

```
http://biodb.jp/hfs.cgi?id=[ID]&type=[ID Type]&db=[Database name]
```

のように [http://biodb.jp/hfs.cgi?]にパラメータを渡すだけで利用できます。設定するパラメータは以下の通りです。

1. [ID]
   変換するIDを指定します。
2. [ID Type]
   指定したIDの形式を指定します。
   利用できるIDのリスト
3. [Database name]
   転送先のデータベース、ビューアーを指定します。
   利用できるデータベースのリスト

それぞれこの画面の下に記載可能なリストがある。

サンプル

Accession Number「BC053657」からH-InvDB (Transcript view)へリンクする場合。

```
http://biodb.jp/hfs.cgi?id=BC053657&type=ACC_ID&db=TRANSCRIPTVIEW
```

HTMLには以下のように記述します。

```
<a href='http://biodb.jp/hfs.cgi?id=BC053657&type=ACC_ID&db=TRANSCRIPTVIEW'>BC053657</a>
サンプル：BC053657
```

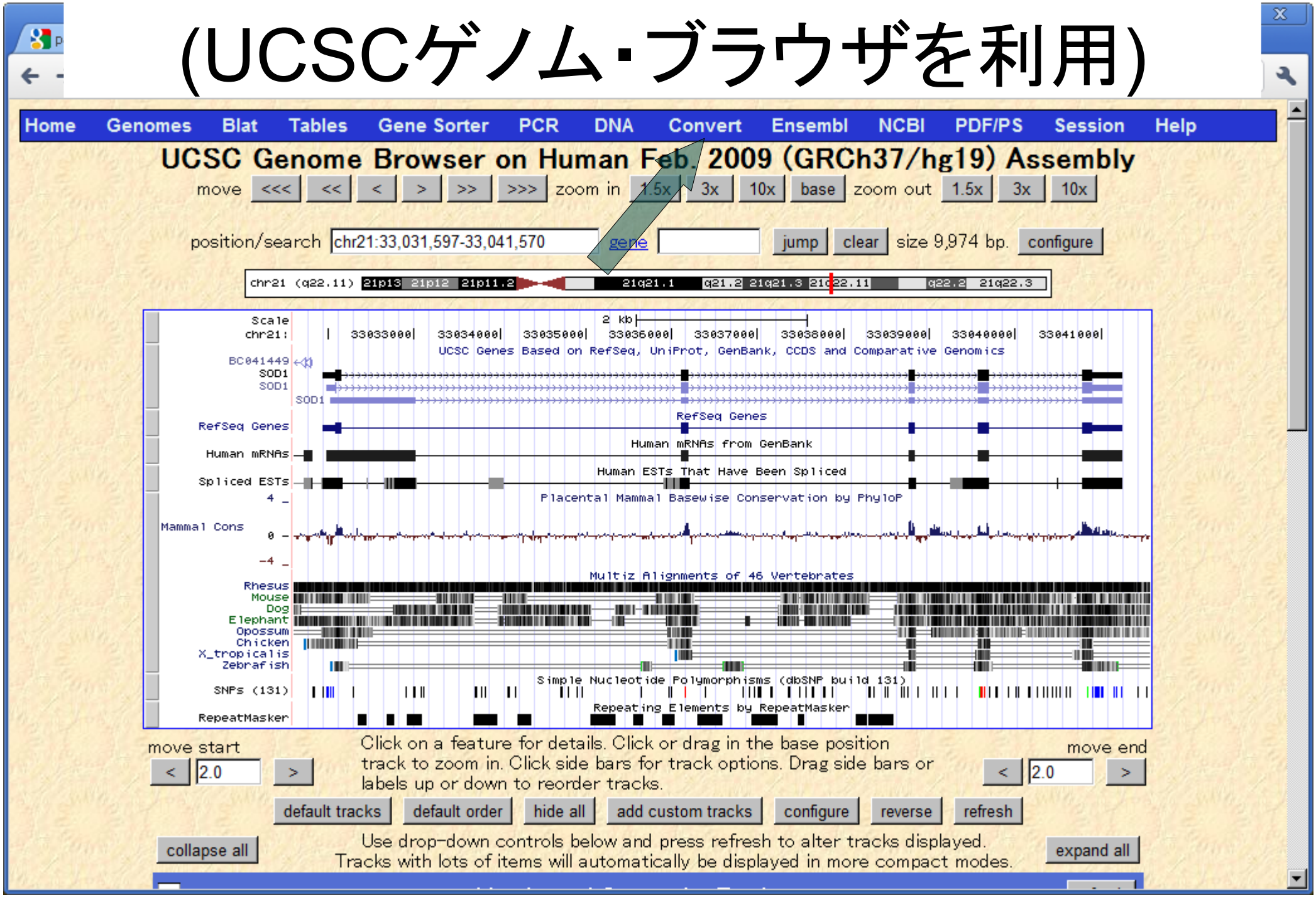
ヒトのID一覧

No.	形式	概要

ゲノム・バージョンの変換
http://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html#Convert

- ゲノムの座標(coordinates)は現在もアッセンブリ（バージョン）の変更に伴い変更されている。
  - scaffold間のgapが埋まる、contigの向きや重複領域の修正などのため。
- UCSCでは3つの変換ツールがある。
  - BLAT
  - Coordinate conversion (ゲノム・ブラウザを利用)
  - Lifting coordinates
    - Web-based
    - Command-line

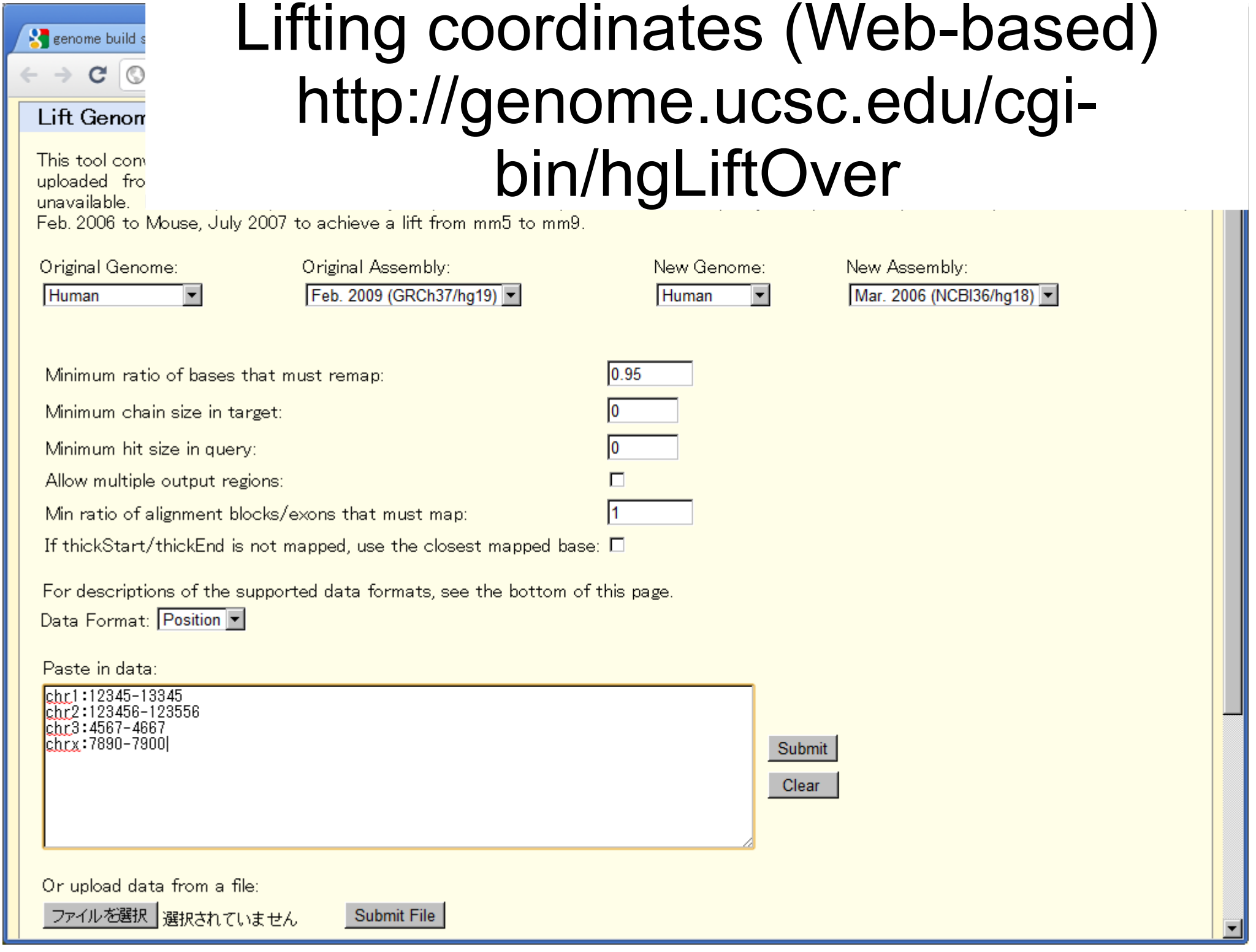
Coordinate conversion (UCSCゲノム・ブラウザを利用)



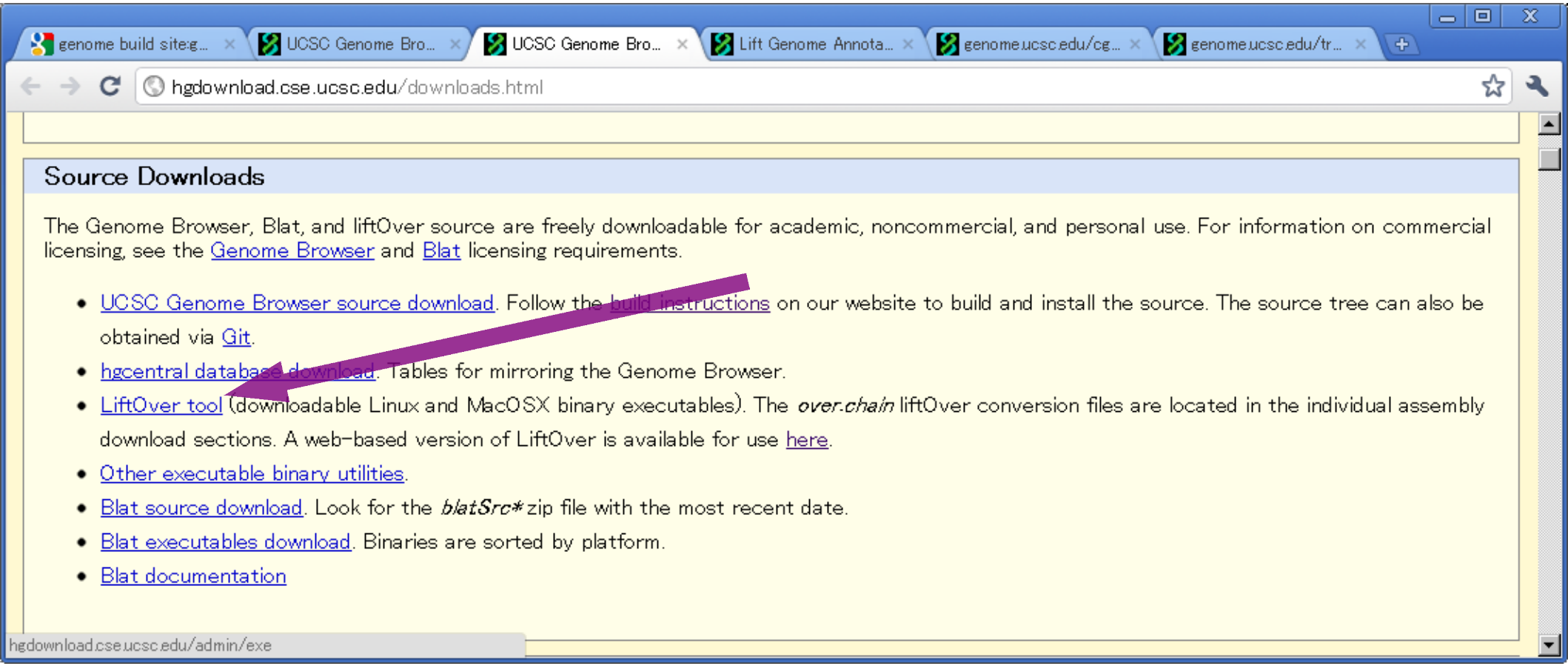
Coordinate conversion

(UCSCゲノム・ブラウザのConvertをクリックしたところ)

Lifting coordinates (Web-based) http://genome.ucsc.edu/cgi-bin/hgLiftOver



Lifting coordinates (command-line)
http://hgdownload.cse.ucsc.edu/downloads.html



アウトライン

- [第1回]
  - 概論
    - ねらい
    - バイオインフォマティクス：どこまで把握すべきか？
  - データベース
- [第2回]
  - ゲノム関連大規模プロジェクト
  - バイオインフォマティクス独学のコツ
  - 生命科学者（実験研究者）にとってのプログラミング
    - アルゴリズム
  - データの形式と変換
    - データ加工変換
    - パース
  - ツール群の分類と見つけ方
    - 配列研究関連ツールの紹介

生命科学解析用ツール群

- ツール選定にあたって
  - 手順
    - 解析手順と各段階での目的の明確化
  - 基準
    - web or local, 人気, 更新年月日
- 選定ガイド
  - 論文から
    - 専門の近い論文 Method
    - NAR web server Issue (http://nar.oxfordjournals.org/content/38/suppl_2)
  - リンク集から

解析ツールリンク集1

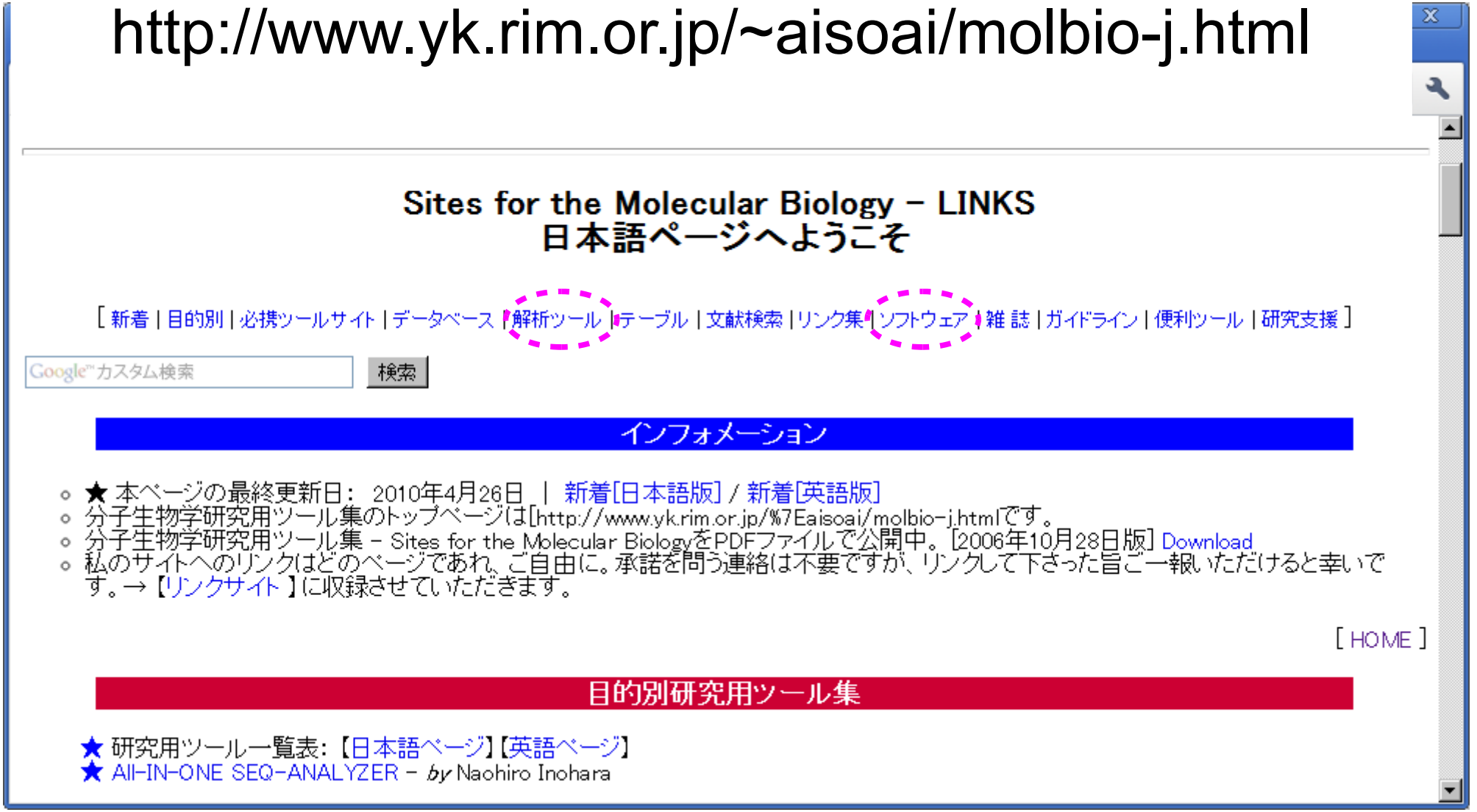
- JST BIRD>ゲノム解析ツール

http://www-btls.jst.go.jp/Links/

解析ツールリンク集2

- 分子生物学研究用ツール集（by Dr. Atsushi Isoai）

http://www.yk.rim.or.jp/~aisoai/molbio-j.html



Sites for the Molecular Biology – LINKS
日本語ページへようこそ

[新着 | 目的別 | 必携ツールサイト | データベース | 解析ツール | テーブル | 文献検索 | リンク集 | ソフトウェア | 雑 誌 | ガイドライン | 便利ツール | 研究支援]

Google™カスタム検索 検索

インフォメーション

- ★ 本ページの最終更新日： 2010年4月26日 | 新着[日本語版] / 新着[英語版]
- 分子生物学研究用ツール集のトップページは[http://www.yk.rim.or.jp/%7Eaisoai/molbio-j.html]です。
- 分子生物学研究用ツール集 – Sites for the Molecular BiologyをPDFファイルで公開中。[2006年10月28日版] Download
- 私のサイトへのリンクはどのページであれ、ご自由に。承諾を問う連絡は不要ですが、リンクして下さった旨ご一報いただけると幸いです。→【リンクサイト 】に収録させていただきます。

[HOME]

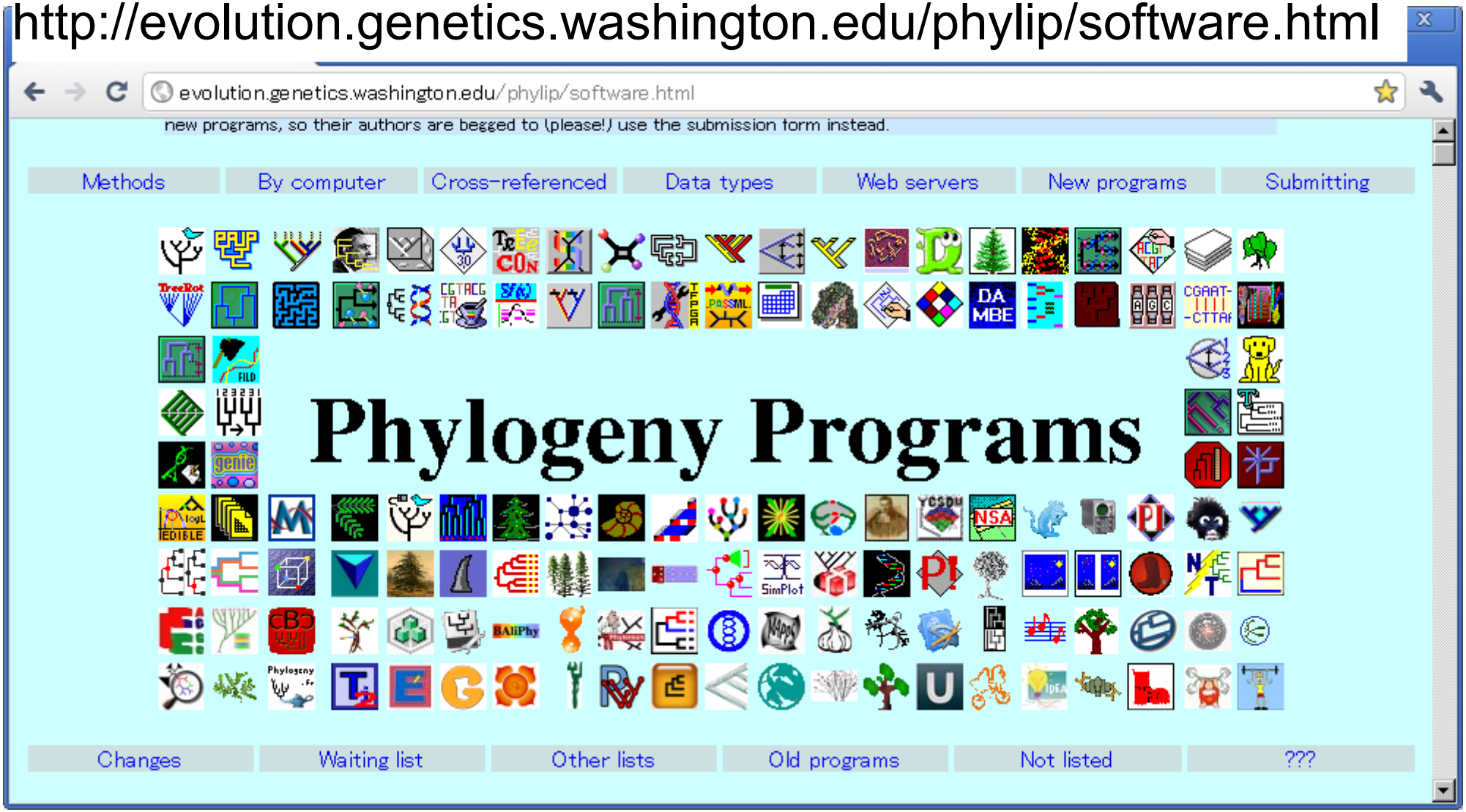
目的別研究用ツール集

★ 研究用ツール一覧表:【日本語ページ】【英語ページ】
★ All-IN-ONE SEQ-ANALYZER – *by* Naohiro Inohara

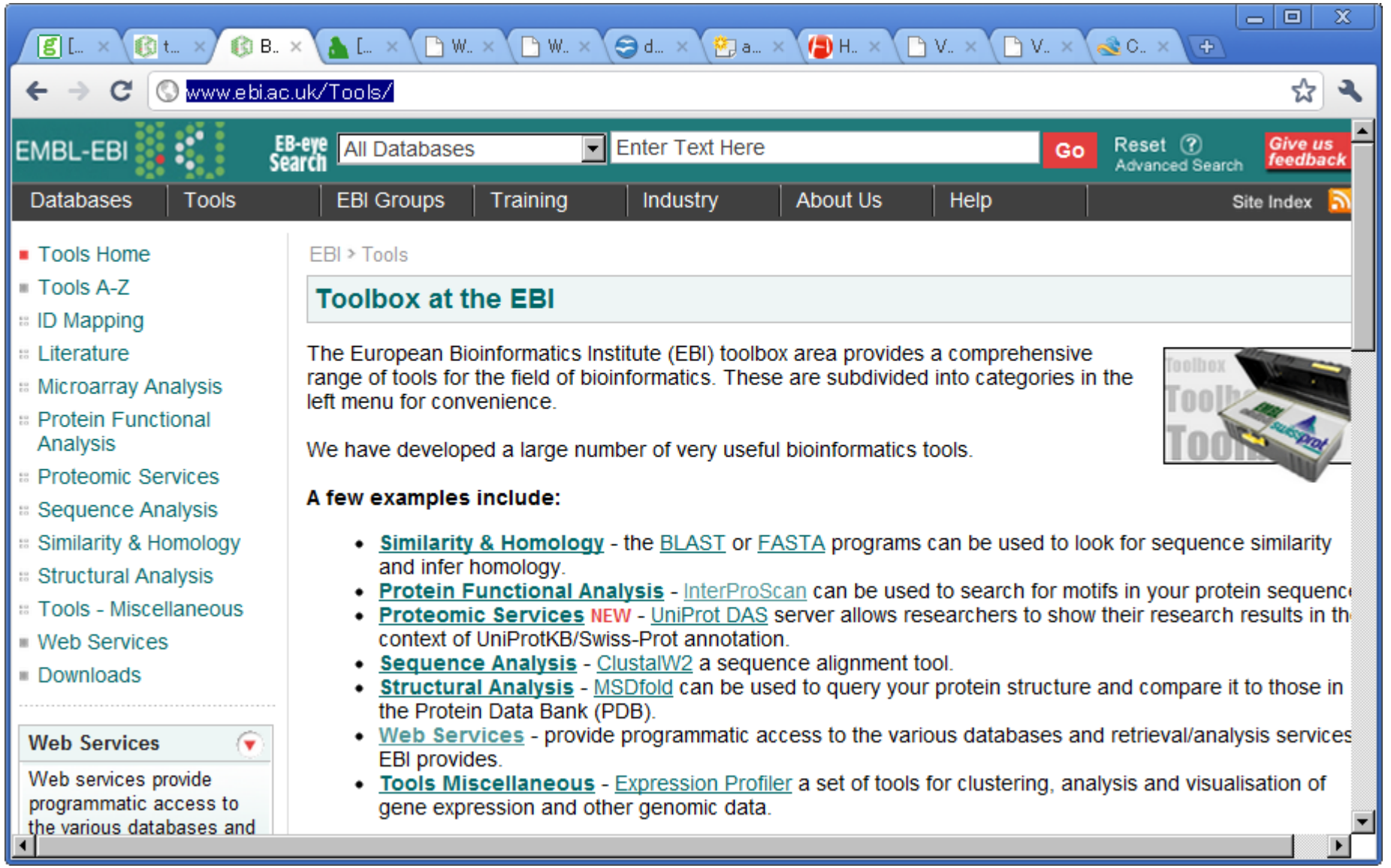
解析ツールリンク集3

- 分子進化学関係 (by Dr.Joe Felsenstein)

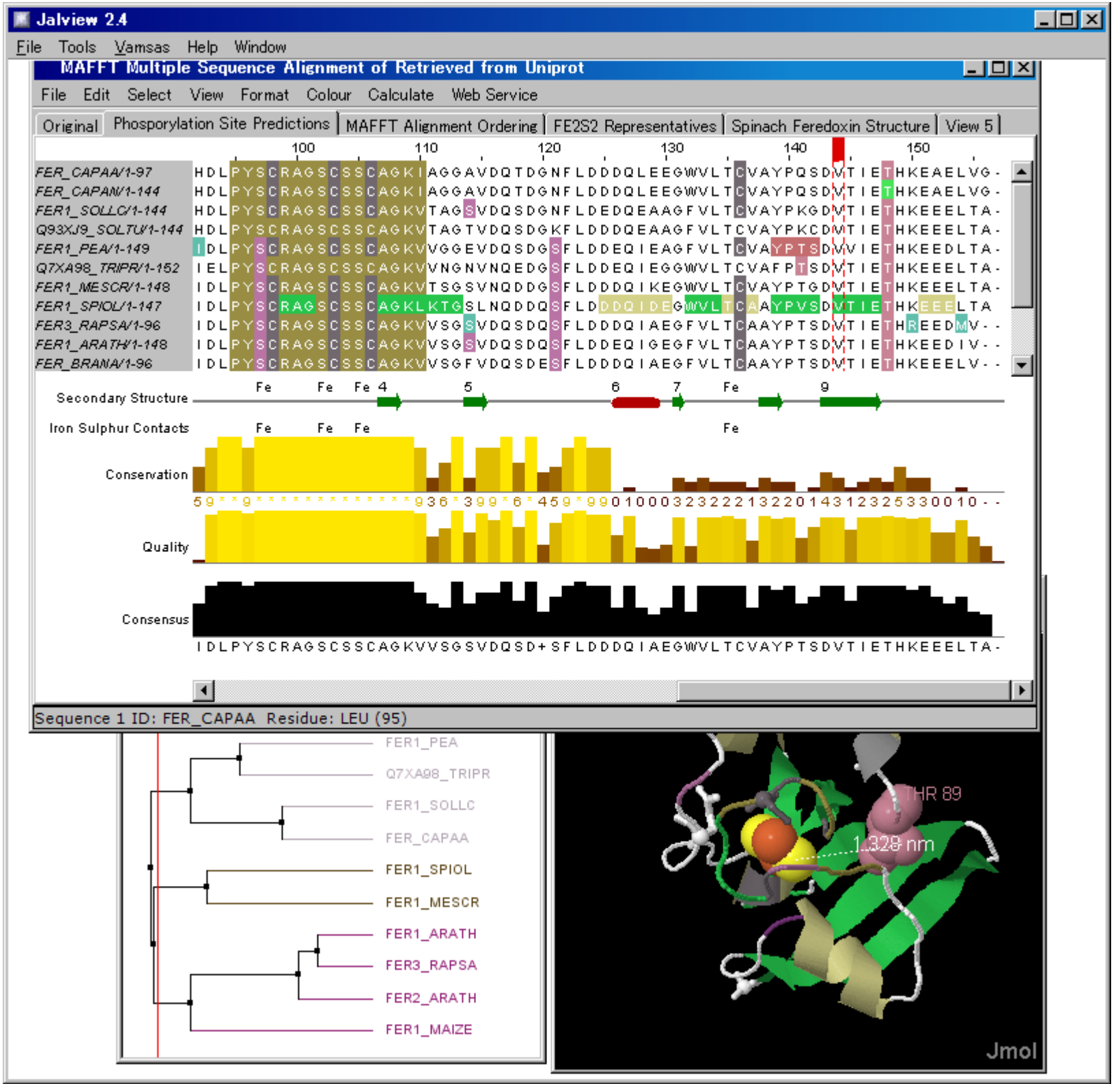
http://evolution.genetics.washington.edu/phylip/software.html



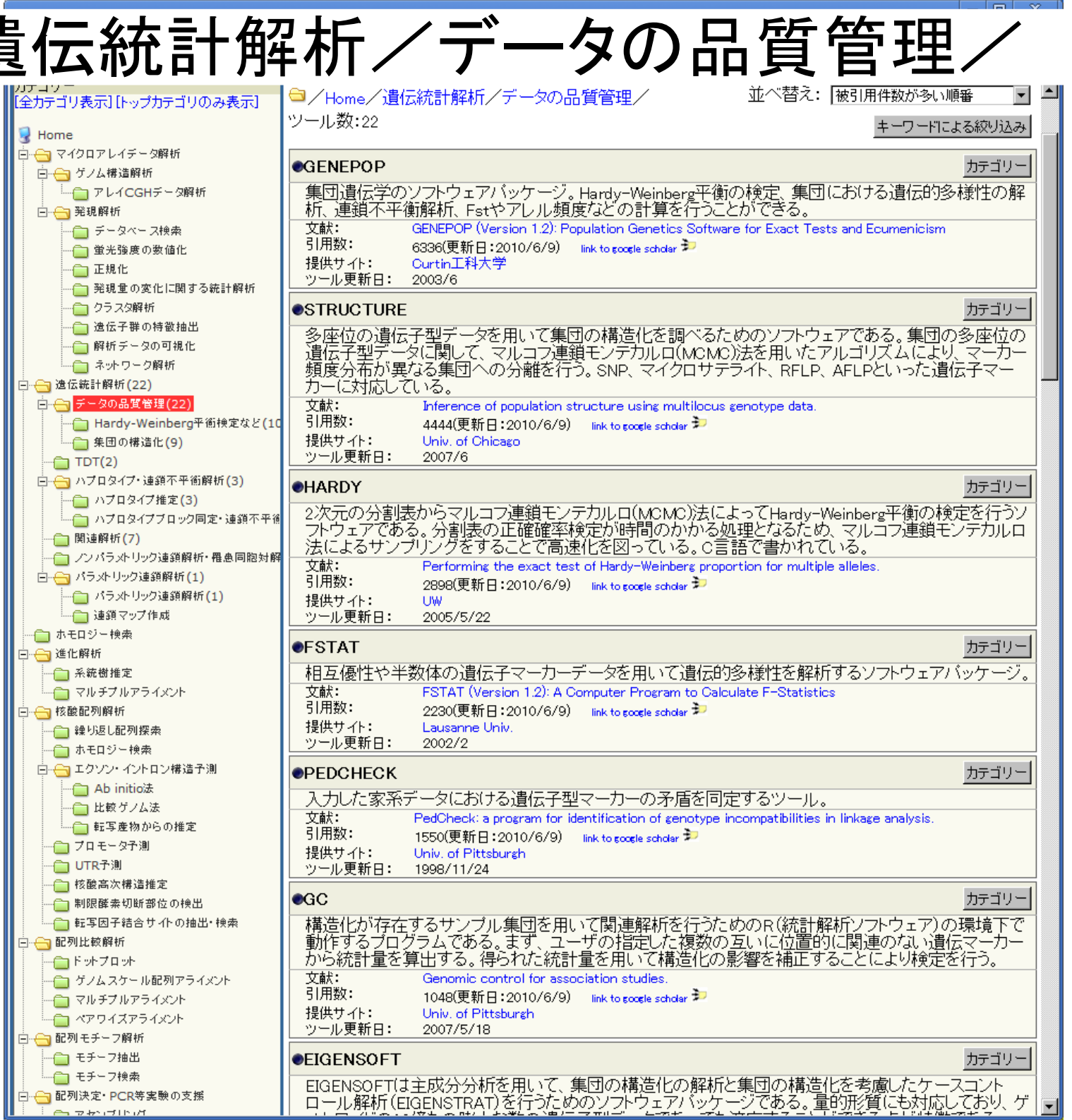
EBI toolbox
http://www.ebi.ac.uk/Tools/



配列editorの例


Jalview 2.4
File Tools Vamsas Help Window
MAFFT Multiple Sequence Alignment of Retrieved from Uniprot
File Edit Select View Format Colour Calculate Web Service
Original
Phosphorylation Site Predictions
MAFFT Alignment Ordering
FE2S2 Representatives
Spinach Feredoxin Structure
View 5
FER_CAPAA/1-97
FER_CAPAN/1-144
FER1_SOLLC/1-144
Q93XJ9_SOLTU/1-144
FER1_PEA/1-149
Q7XA98_TRIPR/1-152
FER1_MESCR/1-148
FER1_SPIOL/1-147
FER3_RAPSA/1-96
FER1_ARATH/1-148
FER_BRANA/1-96
Secondary Structure
Iron Sulphur Contacts
Conservation
Quality
Consensus
Sequence 1 ID: FER_CAPAA Residue: LEU (95)
FER1_PEA
Q7XA98_TRIPR
FER1_SOLLC
FER_CAPAA
FER1_SPIOL
FER1_MESCR
FER1_ARATH
FER3_RAPSA
FER2_ARATH
FER1_MAIZE
THR 89
1.328 nm
Jmol

／Home／遺伝統計解析／データの品質管理／

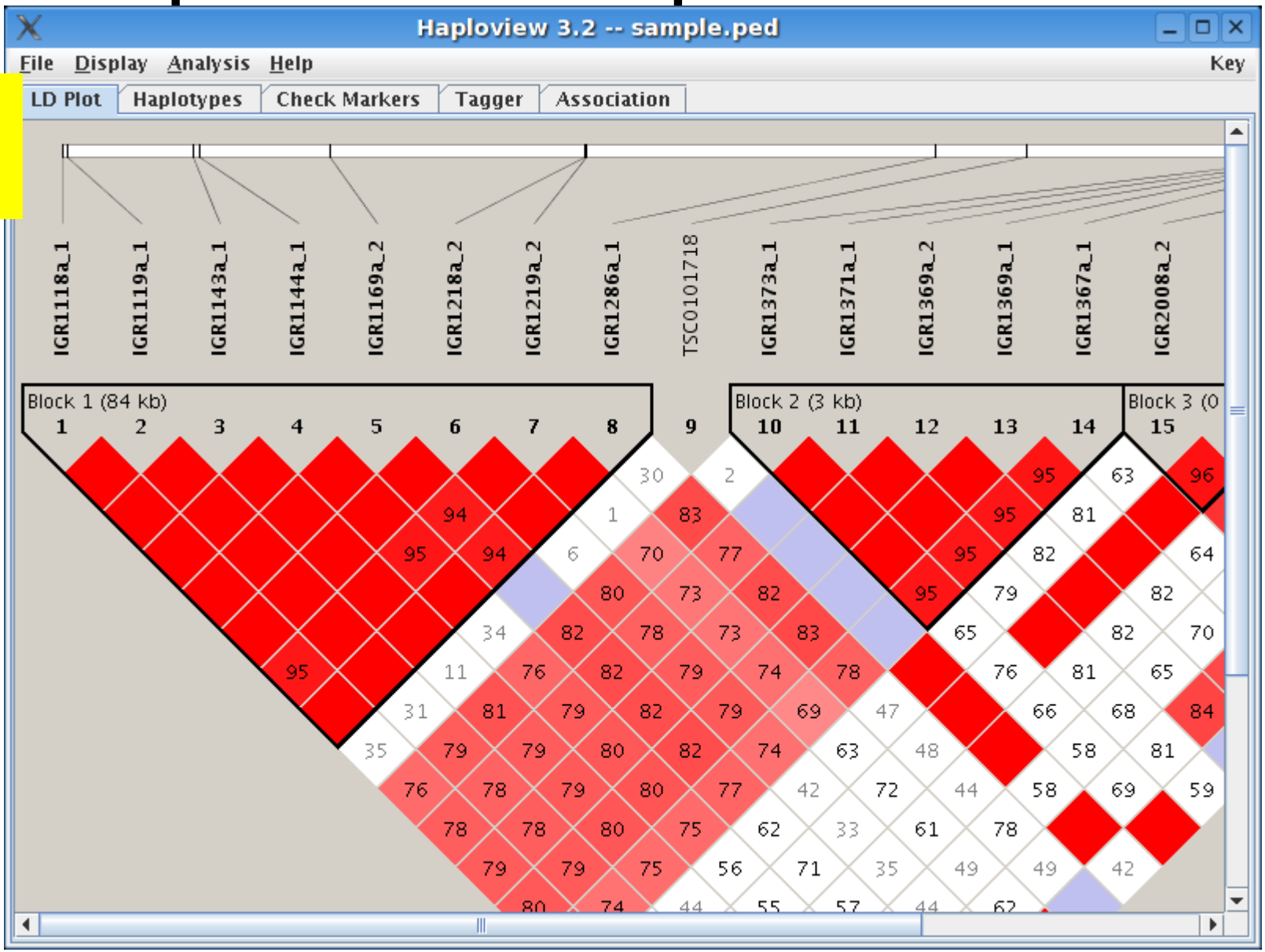


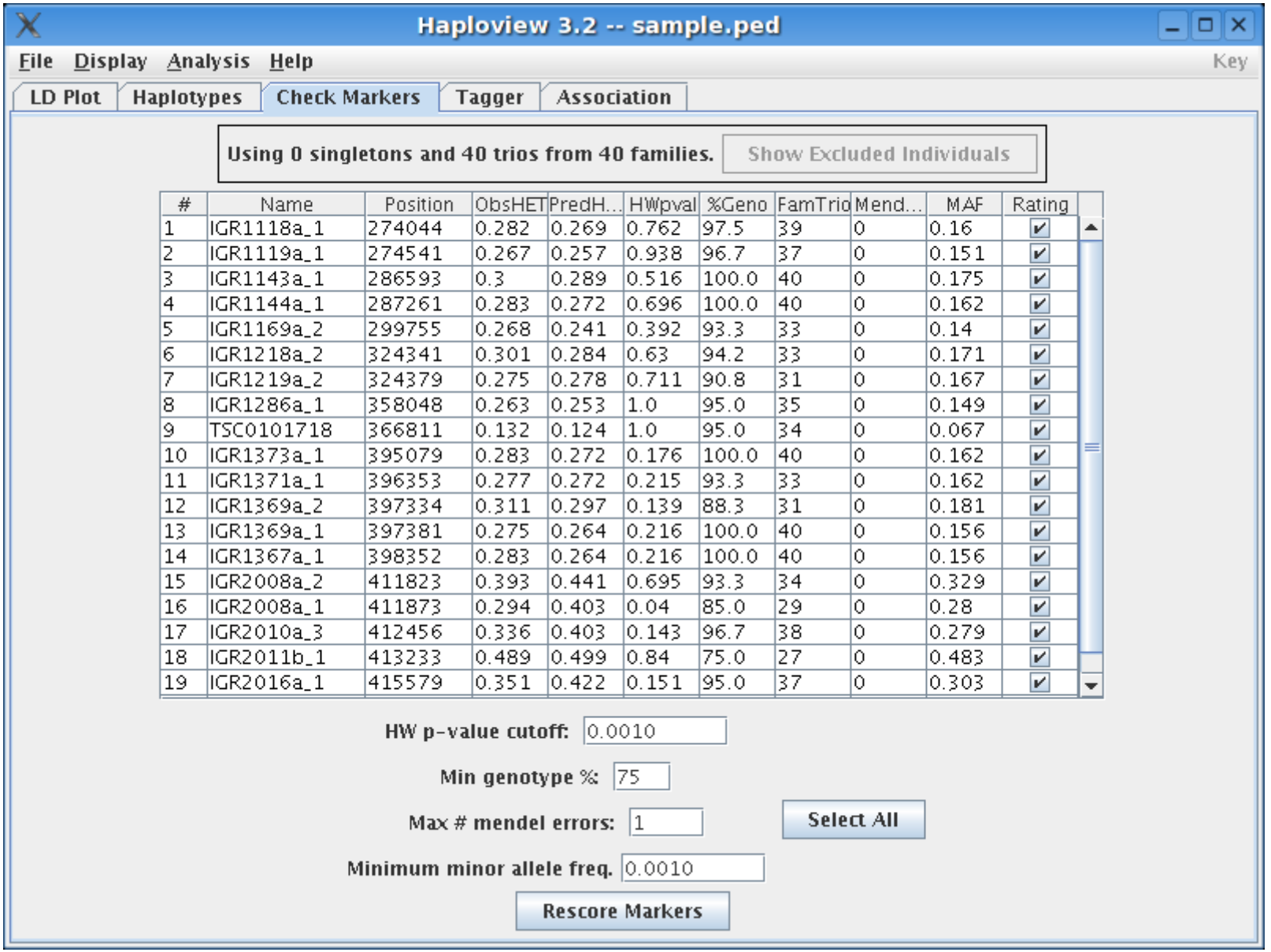
ハプロタイプ解析


ゲノム解析ツール リンク集
User Manual | Broad Instit...
www-btls.jst.go.jp/Links/link.cgi?category=2300
カテゴリー
[全カテゴリ表示] [トップカテゴリのみ表示]
Home
マイクロアレイデータ解析
遺伝統計解析(27)
データの品質管理(3)
Hardy-Weinberg平衡検定など(2)
集団の構造化
TDT
ハプロタイプ・連鎖不平衡解析(27)
ハプロタイプ推定(16)
関連解析(4)
ノンパラメトリック連鎖解析・罹患同胞対解
パラメトリック連鎖解析(1)
パラメトリック連鎖解析(1)
連鎖マップ作成
ホモロジー検索
進化解析
系統樹推定
マルチプルアライメント
核酸配列解析
繰り返し配列探索
ホモロジー検索
エクソン・イントロン構造予測
Ab initio法
比較ゲノム法
転写産物からの推定
プロモータ予測
UTR予測
核酸高次構造推定
制限酵素切断部位の検出
転写因子結合サイトの抽出・検索
配列比較解析
ドットプロット
／Home／遺伝統計解析／ハプロタイプ・連鎖不平衡解析／
並べ替え: 被引用件数が多い順番
ツール数:27
キーワードによる絞り込み
GENEPOP
カテゴリー
集団遺伝学のソフトウェアパッケージ。Hardy-Weinberg平衡の検定、集団における遺伝的多様性の解析、連鎖不平衡解析、Fstやアレル頻度などの計算を行うことができる。
文献: GENEPOP (Version 1.2): Population Genetics Software for Exact Tests and Ecumenicism
引用数: 6336(更新日:2010/6/9) link to google scholar
提供サイト: Curtin工科大学
ツール更新日: 2003/6
HAPLOVIEW
カテゴリー
連鎖不平衡とハプロタイプブロック解析、ブロック内のハプロタイプ推定、SNPあるいはハプロタイプによる関連解析を行い、様々な形式で出力する。HapMapプロジェクトで用いられているタグSNPを選択するアルゴリズムも組み込まれている。また、HapMapプロジェクトが提供している相分離された遺伝子型データをダウンロードする機能がある。Java言語で書かれている。
文献: Haploview: analysis and visualization of LD and haplotype maps.
引用数: 3838(更新日:2010/6/9) link to google scholar
提供サイト: MIT
ツール更新日: 2009/2/27
PHASE
カテゴリー
家系情報のない多数個体の遺伝子型データからマルコフ連鎖モンテカルロ(MCMC)法により集団のハプロタイプ頻度を推定する。ハプロタイプの分布にコアレセンスモデルを仮定しているのが特徴である。実行形式での配布となっている。
文献: A new statistical method for haplotype reconstruction from population data.
引用数: 3118(更新日:2010/6/9) link to google scholar
提供サイト: UW
ツール更新日: Jun-04
haplo.stats
カテゴリー
haplo.stats(旧称haplo.score)は多型の遺伝子型データからEMアルゴリズムを使用してハプロタイプ推定を行うS-PLUS/R(統計解析ソフトウェア)の環境下で動作するプログラム群である。ハプロタイプ推定の結果を用いて、関心のある表現型との関連を検定することができる。
文献: Score tests for association between traits and haplotypes when linkage phase is ambiguous.
引用数: 1141(更新日:2010/6/9) link to google scholar
提供サイト: Mayo Clinic

Haploview: LD plot

連鎖不平衡
ハプロタイプブロック
タグSNP




Haploview 3.2 -- sample.ped
File Display Analysis Help
Key
LD Plot Haplotypes Check Markers Tagger Association
Using 0 singletons and 40 trios from 40 families.
Show Excluded Individuals
Name Position ObsHET PredH... HWpval %Geno FamTrio Mend... MAF Rating
1 IGR1118a_1 274044 0.282 0.269 0.762 97.5 39 0 0.16
2 IGR1119a_1 274541 0.267 0.257 0.938 96.7 37 0 0.151
3 IGR1143a_1 286593 0.3 0.289 0.516 100.0 40 0 0.175
4 IGR1144a_1 287261 0.283 0.272 0.696 100.0 40 0 0.162
5 IGR1169a_2 299755 0.268 0.241 0.392 93.3 33 0 0.14
6 IGR1218a_2 324341 0.301 0.284 0.63 94.2 33 0 0.171
7 IGR1219a_2 324379 0.275 0.278 0.711 90.8 31 0 0.167
8 IGR1286a_1 358048 0.263 0.253 1.0 95.0 35 0 0.149
9 TSC0101718 366811 0.132 0.124 1.0 95.0 34 0 0.067
10 IGR1373a_1 395079 0.283 0.272 0.176 100.0 40 0 0.162
11 IGR1371a_1 396353 0.277 0.272 0.215 93.3 33 0 0.162
12 IGR1369a_2 397334 0.311 0.297 0.139 88.3 31 0 0.181
13 IGR1369a_1 397381 0.275 0.264 0.216 100.0 40 0 0.156
14 IGR1367a_1 398352 0.283 0.264 0.216 100.0 40 0 0.156
15 IGR2008a_2 411823 0.393 0.441 0.695 93.3 34 0 0.329
16 IGR2008a_1 411873 0.294 0.403 0.04 85.0 29 0 0.28
17 IGR2010a_3 412456 0.336 0.403 0.143 96.7 38 0 0.279
18 IGR2011b_1 413233 0.489 0.499 0.84 75.0 27 0 0.483
19 IGR2016a_1 415579 0.351 0.422 0.151 95.0 37 0 0.303
HW p-value cutoff: 0.0010
Min genotype %: 75
Max # mendel errors: 1
Select All
Minimum minor allele freq. 0.0010
Rescore Markers

Genepop

Genepop on the Web

GENEPOP is a population genetics software package originally developed by *Michel Raymond* (*Raymond@isem.univ-montp2.fr*) and *Francois Rousset* (*Rousset@isem.univ-montp2.fr*), at the Laboratiore de Genetique et Environment, Montpellier, France. The latest version of Genepop (4.0) is now available from http://kimura.univ-montp2.fr/~rousset/Genepop.htm. Genepop 4.0 runs under Windows, and can also be compiled to run under Unix or Linux. It will compile on Mac OSX machines if you have the developer tools installed. To compile under Unix or Linux, open a terminal window and cd to the Genepop source directory. Then issue the command:

'g++ -DNO_MODULES -o Genepop GenepopS.cpp -O3'

This latest version is easier to use and has some additional analyses (compared to v3.4) plus the ability to run in Batch mode.

The web version is still available for teaching purposes and for those who, for some reason, cannot run the latest version on their local PC or Mac. Below is the Genepop WWW menu with links to the data input and help pages. For further information on the Genepop program and its web implementation see the history page.

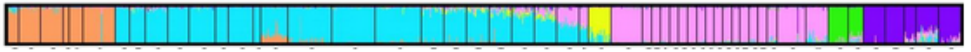
Option	Status of Web Version	Help Files
1. Hardy Weinberg Exact Tests	Upgraded to Genepop 4.0.10 (compiled binary from source code provided by Francois Rousset)	Option 1 Help
2. Linkage Disequilibrium	Upgraded to Genepop 4.0.10 (compiled binary from source code provided by Francois Rousset)	Option 2 Help
3. Population Differentiation	Upgraded to Genepop 4.0.10 (compiled binary from source code provided by Francois Rousset)	Option 3 Help
4. Nm estimates	Upgraded to Genepop 4.0.10 (compiled binary from source code provided by Francois Rousset)	Option 4 Help
5. Basic Information, Fis and gene diversities	Upgraded to Genepop 4.0.10 (compiled binary from source code provided by Francois Rousset)	Option 5 Help
6. Fst & other correlations	Upgraded to Genepop 4.0.10 (compiled binary from source code provided by Francois Rousset)	Option 6 Help
7. File Conversion	Equivalent to Dos versions 3.4. Includes additional file conversion to ARLEQUIN format.	Option 7 Help
8. Miscellaneous Utilities	Upgraded to Genepop 4.0.10 (compiled binary from source code provided by Francois Rousset)	Option 8 Help

Additional Help Files

- Data input format
- Appendix 1 (null allele estimates, exact tests, markov chain probabilities, test statistics)
- Appendix 2 (Multilocus F-statistics)
- Appendix 3 (Microsatellite allele sizes, R_{ST}, and ρ_{ST}, Robertson and Hill's estimator of F_{IS}, Bootstraps)
- Bibliography

Structure

The program *structure* is a free software package for using multi-locus genotype data to investigate population structure. Its uses include inferring the presence of distinct populations, assigning individuals to populations, studying hybrid zones, identifying migrants and admixed individuals, and estimating population allele frequencies in situations where many individuals are migrants or admixed. It can be applied to most of the commonly-used genetic markers, including SNPS, microsatellites, RFLPs and AFLPs.



Download Structure 2.3.3.

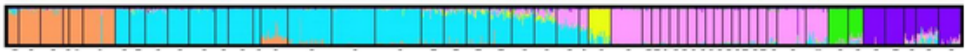
What to cite: The basic algorithm was described by Pritchard, Stephens & Donnelly (2000). Extensions to the method were published by Falush, Stephens and Pritchard (2003) and (2007) and by Hubisz, Falush, Stephens and Pritchard (2009).

Contributors: Daniel Falush, Melissa Hubisz, Matthew Stephens, Jonathan Pritchard, Peter Donnelly, William Wen, Mike Trienis, Pall Melsted.

Questions and Discussion: We have now started a Google Groups forum devoted to Structure. This replaces the Genetic Software Forum which is no longer active.

Plotting programs and other resources: ***CLUMPP*** and ***distruct*** from Noah Rosenberg's lab can automatically sort the cluster labels and produce nice **graphical displays** of *structure* results. Other plots are produced directly by the software package itself. A **free publicly available cluster** has kindly been made available for running computationally intensive *structure* jobs by CBSU at Cornell. Xavier Didelot's program xmfa2struct **converts files in eXtended Multi-Fasta (XMFA)** format into Structure input format.

Sample data sets: available here.

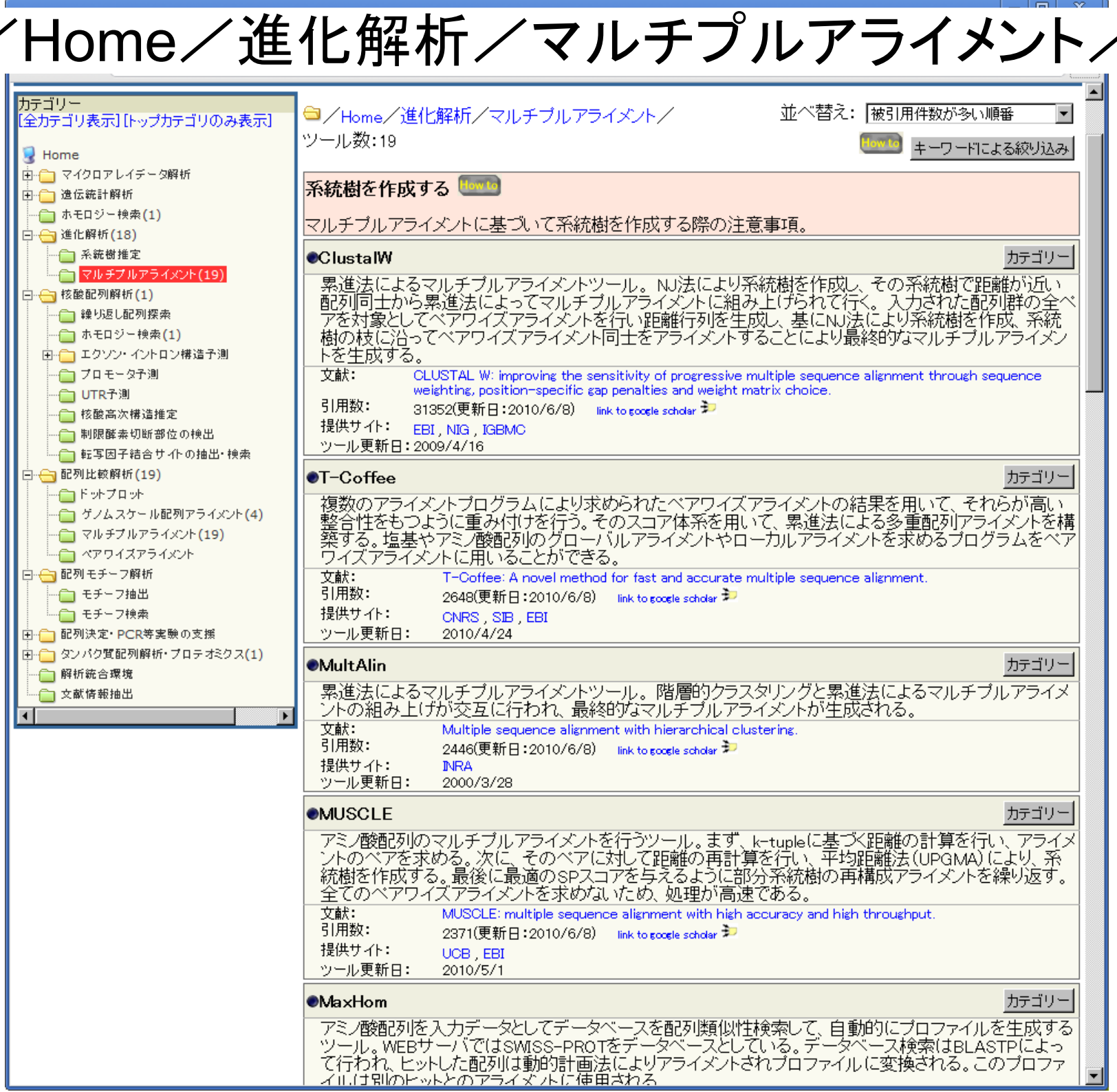


Taita thrush: An example of MCMC convergence based on the original paper is shown here.

Some miscellaneous applications: *structure* has been widely used for interpreting population structure of humans and other organisms. A selection of interesting references (mainly applications) is shown below.

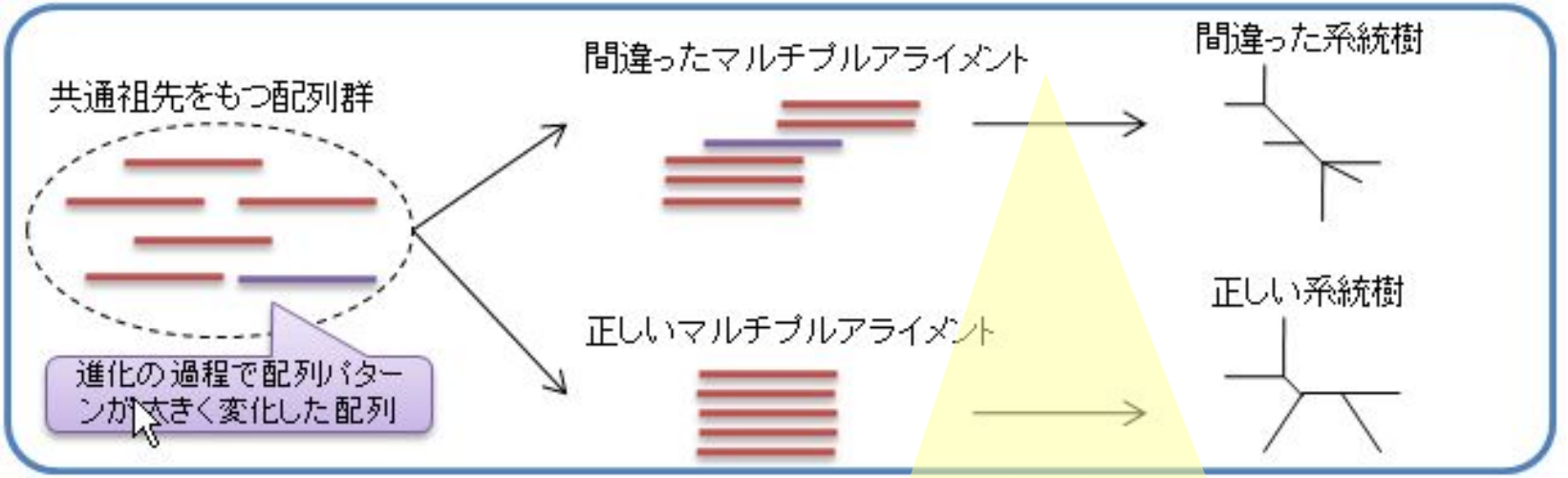
Traces of human migrations in Helicobacter pylori populations. D. Falush, T. Wirth,

／Home／進化解析／マルチプルアライメント／



マルチプルアライメントに基づいて系統樹を作成する際の注意事項1

- 『系統樹を作成する』
  - 【マルチプルアライメントの精度と系統樹】



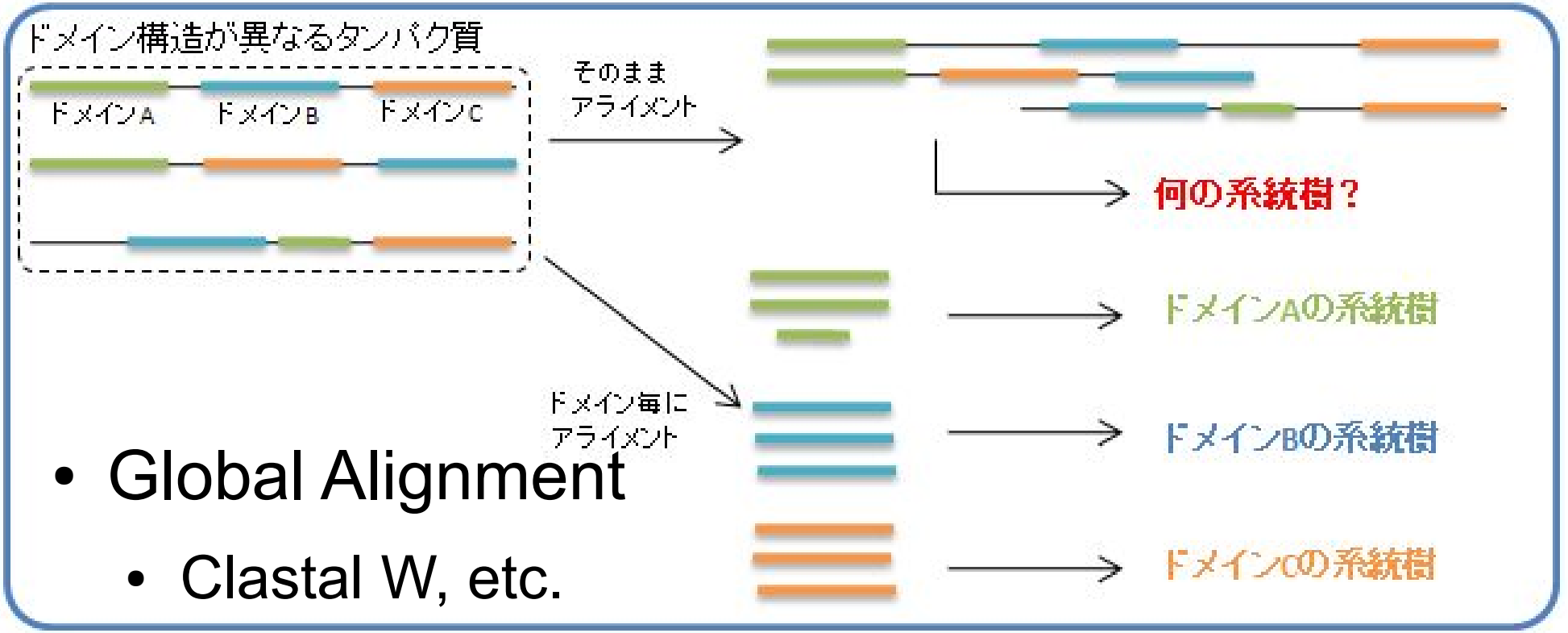
  - 【マルチドメインタンパク質のマルチプルアライメント】

配列エディターでギャップ位置を手直し、保存。
さらにマルチプルアライメントソフトにかける。

http://www-btls.jst.go.jp/Links/howto.cgi?id=4000_01

マルチプルアライメントに基づいて系統樹を作成する際の注意事項

【マルチドメインタンパク質のマルチプルアライメント】



- Global Alignment
  - Clastal W, etc.
- Local Alignment
  - FASTA, BLAST, etc.

http://fasta.bioch.virginia.edu/fasta_www2/fasta_list2.shtml

FASTA Sequence Comparison at the U. of Virginia

UVa FASTA Server

About
- Getting started

Other FASTA Servers
- EMBL-EBI
- KEGG (Japan)

References
- FASTA
- FASTX/FASTY
- Statistics
- FASTS/FASTF

Software
- FASTA v36 ChangeLog
- Downloads
- Sequence Libraries
- Developer Mailing list

Other resources
- CHAPS - Convert HMMs and Profiles
- Near optimal alignments
- FASTA Exercises
- NCBI BLAST server
- EMBL-EBI Server

The **FASTA** programs find regions of local *or global (new)* similarity between Protein or DNA sequences, either by searching Protein or DNA databases, or by identifying local duplications within a sequence. Other programs provide information on the statistical significance of an alignment. Like **BLAST**, **FASTA** can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

Protein
- Protein-protein **FASTA**
- Protein-protein Smith-Waterman (**ssearch**)
- (New) Global Protein-protein (Needleman-Wunsch) (**ggsearch**)
- (New) Global/Local protein-protein (**glsearch**)
- Protein-protein with unordered peptides (**fasts**)
- Protein-protein with mixed peptide sequences (**fastf**)

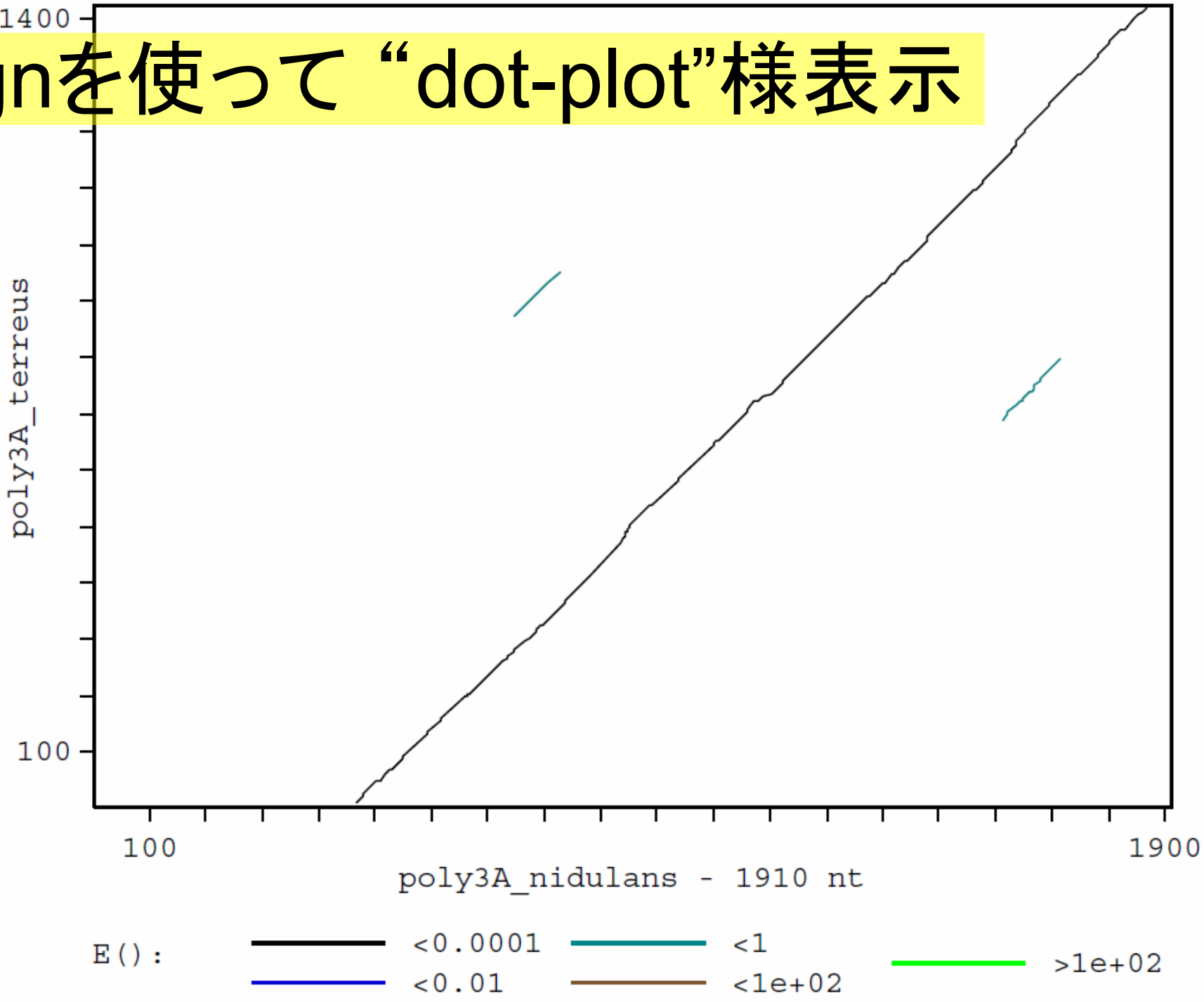
Nucleotide
- Nucleotide-Nucleotide (DNA/RNA **fasta**)
- Ordered Nucleotides vs Nucleotide (**fastm**)
- Un-ordered Nucleotides vs Nucleotide (**fasts**)

Translated
- Translated DNA (with frameshifts, e.g. ESTs) vs Proteins (**fastx/fasty**)
- Protein vs Translated DNA (with frameshifts) (**tfastx/tfasty**)
- Peptides vs Translated DNA (**tfasts**)

Statistical Significance
- Protein vs Protein shuffle (**prss**)
- DNA vs DNA shuffle (**prss**)
- Translated DNA vs Protein shuffle (**prfx**)

Local Duplications
- Local Protein alignments (**lalign**)
- Plot Protein alignment "dot-plot" (**plalign**)
- Local DNA alignments (**lalign**)
- Plot DNA alignment "dot-plot" (**plalign**)

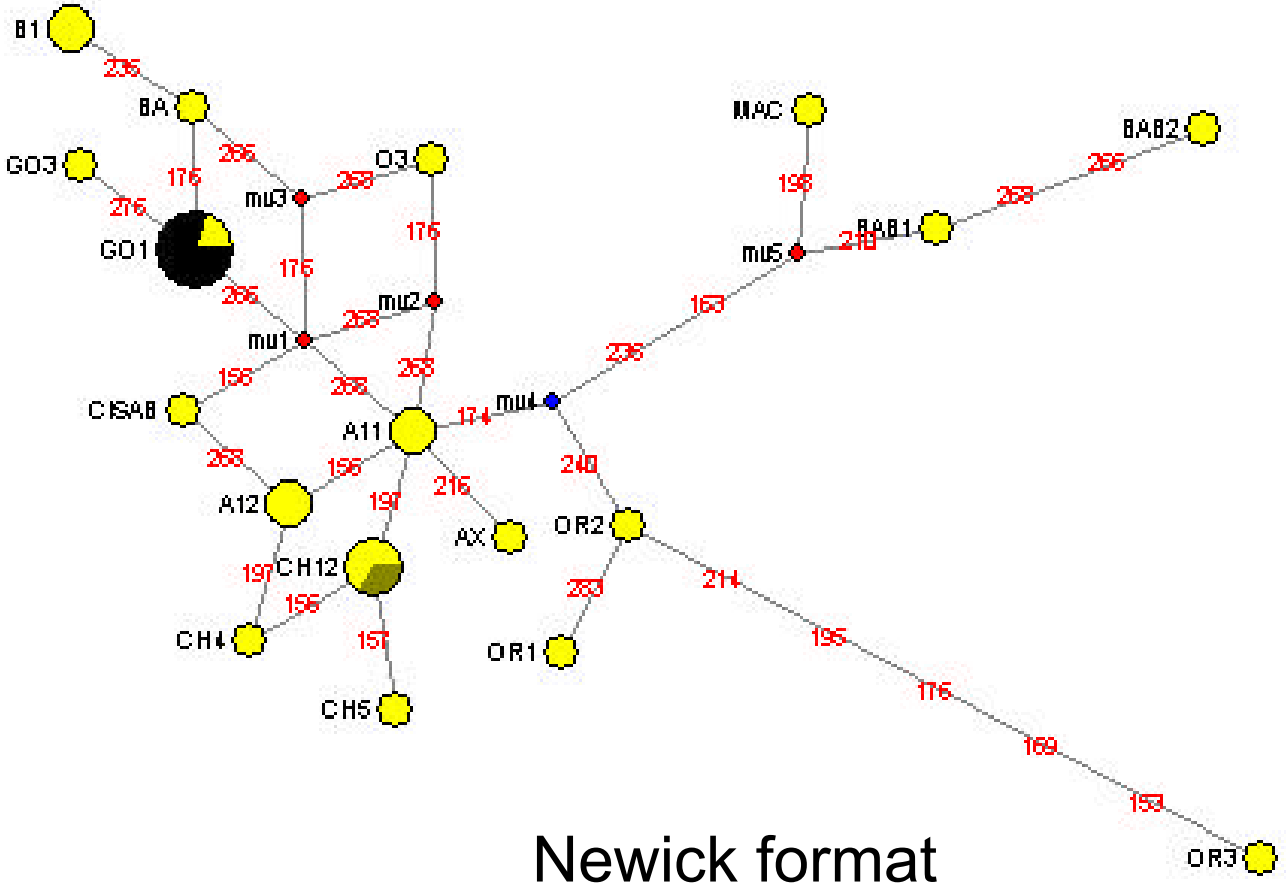
palignを使って “dot-plot”様表示



／Home／進化解析／系統樹推定／

- 系統樹 phylogenetic tree
  - 距離法
    - UPGMA法
    - NJ法
  - 最節約法
  - 最尤法
  - ベイズ法
- 系統ネットワーク phylogenetic network
  - Network 4.516
  - http://www.fluxus-engineering.com/sharenet.htm

系統樹・系統ネットワーク関連用語



node, cluster

edge, blanch

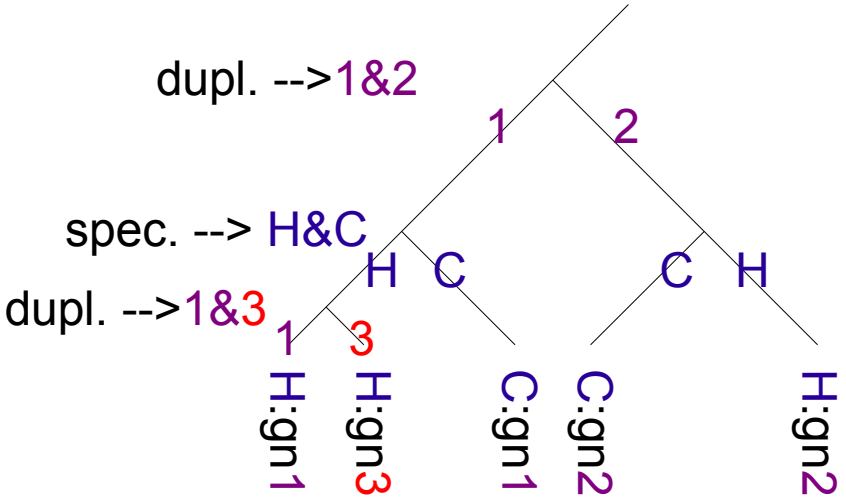
leaf node, OTU

Newick format

(MAC(RAB1, BAB2))mu5, ((OR1,OR3)OR2)mu4

((MAC:1(RAB1:0, BAB2:2):1)mu5:2, ((OR1:1,OR3:5)OR2:1))mu4

相同遺伝子

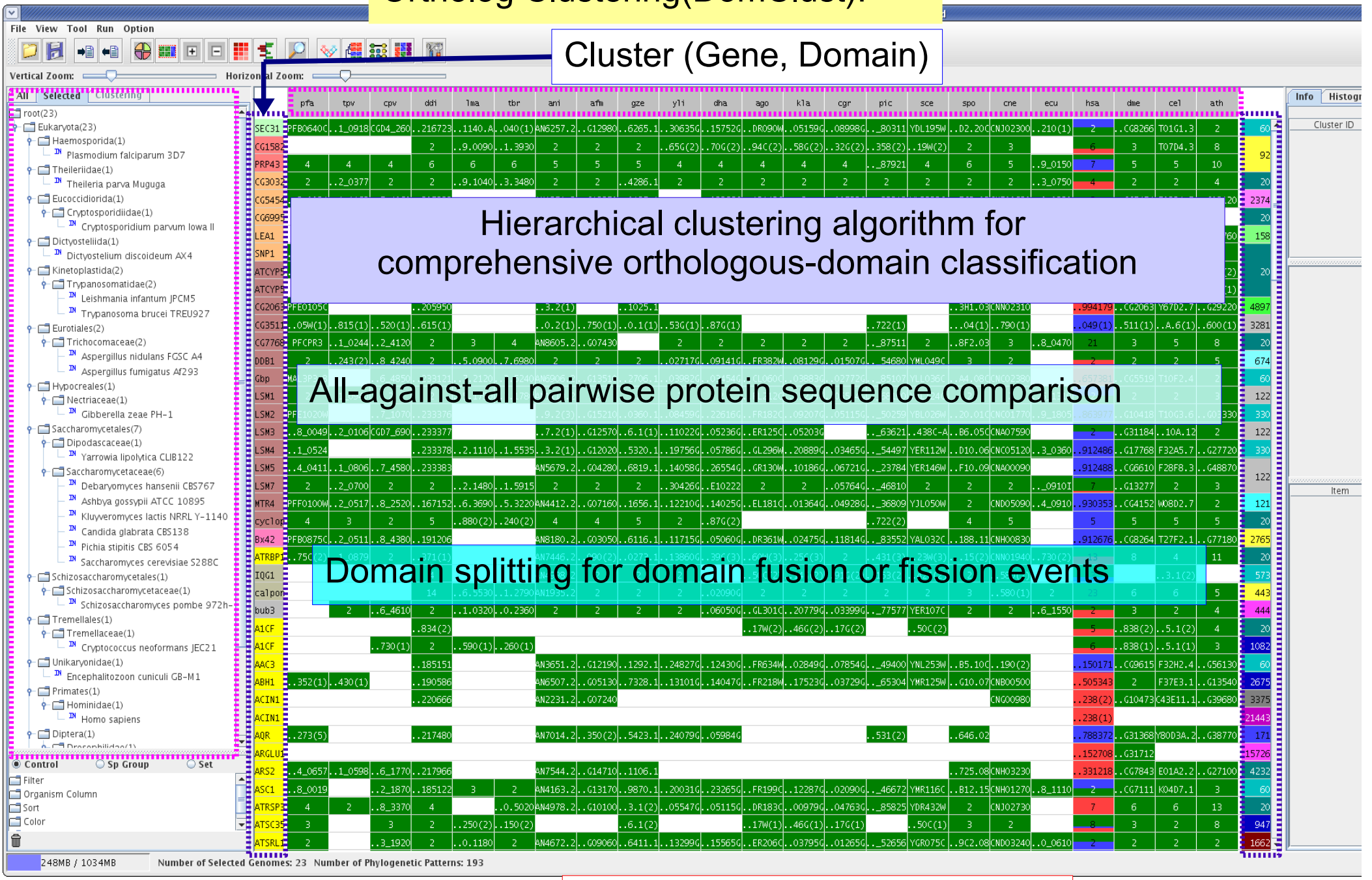


- homology
  - Orthology
    - 種分化(speciation)によって分かれた
  - Paralogy
    - 遺伝子重複 (gene duplication)によって分かれた
  - Ohnology
    - 全ゲノム重複 (whole-genome duplication)によって分かれた

＊遺伝子の機能による分類でないことに注意

RECOG (Research Environment for Comparative Genomics)

Ortholog Clustering(DomClust):



http://mbgd.nibb.ac.jp/RECOG/