

よくわかるバイオインフォマティクス入門』まとめノート

@cakkby2 (自分用まとめ)

§1. 配列解析

4/22

1.1 配列解析のキヤ

- DNAに生じる変異で配列解析上重要なものの：点変異、挿入変異、欠失変異
→ 集団内で変異型が固定され、全てを占めた状態 = 置換

・**負の選択** = 集団内から生存に不利な変異が排除されること

・**正の選択** = 集団内に生存に有利な変異が広まること

中立なものは **遺伝的浮動** である

- 共通祖先から分歧した 酸基/3ミ酸 配列 ⇒ 相同

オーソロガス：種分化に伴い分歧した相同配列間の関係

パラロガス：遺伝子重複により生じた相同な配列間の関係

1.2 配列データベースと配列検索

- DNA配列のDB

NCBIのGenBank, EMBL-EBIのENA, NIGのDDBJ

- 3ミ酸配列のDB → EMBL-EBIのUniProt

- データベース検索：キーワード検索・配列類似性検索

- 配列データの形式は FASTA が多い

ex. アクセサリ蛋白名 遺伝子名 種名
→ AAA66(66.1) alpha A-crystallin [Nannopalaehrenbergi]
MDVTTIQ... ← 配列のN末端から(5'端)

NCBI BLAST で！

Basic Local Alignment
Search Tool

1.3 配列アライメント

BLAST

- 大域的アライメントと局所的アライメント

- 多重配列アライメント (マルチアルゴリズム)

⇒ ツリーベース法：まずペアワイズアライメントして仮の系統樹を作り(木構造).これに基づいて重ね合わせ...

- 鹿猿など直しにくい欠点

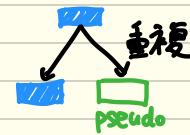
ソフトウェア: MAFFT, T-Coffee, MUSCLE, ClustalW2 ...

§2. 分子系統解析

4/23

2.2 遺伝子多様化の分子機構

- **遺伝子重複**: 不等交差や染色体・ゲムの倍加による。機能を失ったものの(やncRNAの調節部など)は**偽遺伝子**と呼ぶ。



- **ドメインシャッピング**: タンパクドメインの組み合わせ

- **遺伝子変換**: 重複遺伝子間での塩基配列の交換



- **遺伝子水平伝播**: 生物種を超えた遺伝子の伝達。ウイルスなどで取り込まれるもの...

2.3 分子進化学的解析

- まずは配列3ラインメントで相違度を計算

- 同じ座位置の置換(多重置換), 分岐後に独立して置換(水平置換)
↑元に戻ることも(復帰置換)



- **進化距離の推定**

Jukes-Cantor モデル, K80 モデル... 最尤法で
修正には ガウス補正, Poisson 表記などを

- **(非)同義置換の推定**

同義での進化速度 $r_s >$ 非同義での進化速度 r_A ω 比 $\frac{r_A}{r_s} < 1 \xrightarrow{\text{はる}} \approx 1$ (はず)

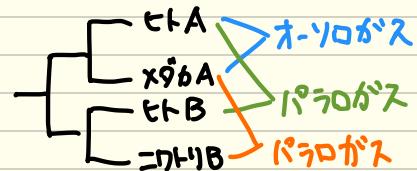
ω 比が 1 より大きい \Rightarrow 正の選択圧が働いたと考える

- **分子系統樹**

OTU (Operational Taxonomic Unit): 分子系統樹上で扱われる生物種・遺伝子
樹形分類

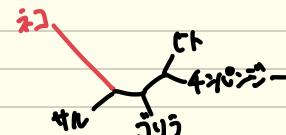
$$\text{無根: } t = (2n-5)! / [2^{n-3}(n-3)!]$$

$$\text{有根: } t = (2n-3)! / [2^{n-2}(n-2)!]$$



- **分子進化速度の一定性**を仮定できないときは、無根系統樹のみ推定可能

\rightarrow OTUにおける外群群の設定



- **系統樹の推定**

UPGMA?

距離並行法: 平均一致法・Fitch-Margoliash 法・最小進化法・最小二乗法・近隣結合法(NJ法)

形質状態法: 最節約法・最尤法・ベイズ法

△部分木の信頼性
はブートストラップ確率

\Rightarrow 近年たくさん得られたゲノム情報を活かす!
ベイズ法・最尤法・NJ法を比較・検討...

§3. タンパク質の立体構造解析

4/27

3.1 立体構造の成り立ち

・立体構造DB → **wwPDB** (worldwide Protein Data Base)

・PDBに登録されているフォーマット 見るソフト: UCSF Chimera, PyMOL, VMD ...

旧 PDB フォーマット 各行 80文字の固定幅!

HEADER	OXYGEN STORAGE/TRANSPORT	05-NOV-98	IBZ1					
TITLE	HEMOGLLOBIN (ALPHA + MET) VARIANT							
ATOM	1 N MET A 1 15.774 28.408 41.946 1.00 69.06 N							
ATOM	2 CA MET A 1 19.105 28.442 42.578 1.00 83.35 C							
:	:	:	:					
HETATM	4405 FE HEM A 143 18.364 18.424 23.704 1.00 13.78 FE							
:	:	:	:					
番号	原子名	残基名	鎖	残基	座標 [Å]	傾斜	温度因子	元素

mmCIF フォーマット

最初は キーバリュー形式で 項目と値をペアで表示

その後は 表形式で 項目を左端に 値をスペース区切りで、値の長さは可変

3.3 立体構造の比較と分類

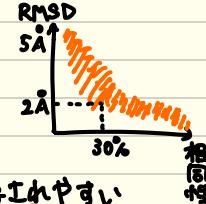
平行βシートが

・主なタンパクの構造クラス: All-alpha, All-beta, alpha/beta, alpha+beta
主な
主でない

→これらの組み合わせ: 構造モチーフ = 超二次構造 βターンやβ-α-βなど

・(通常の)配列相同性で発見にくい遠縁の相同タンパクのグループ: スーパーファミリー

配列相同性が 30% 以下でも RMSD はおよそ 2 Å 以下で 立体構造は進化的に保存されやすい



・ファミリーやフォールドのDB: SCOP, CATH ... ← 特定のフォールド (super-fold) を含む

・立体構造の(構造類似性による)アライメント ⇒ 効率的なアルゴリズムは存在しない (証明済み)

↑ DPは使えない (スコア関数が2体以上)

⇒ 近似解 (DALI, CE, MATRAS などや PyMOLなど)

3.4 立体構造予測

・de novo 法 - 1. 分子力学的アプローチ 2. 統計力学的アプローチ

局所構造を導入できる

1. 分子力学的: ポテンシャルの導入 粗視化モデルなどを連続最適化

↳ ROSETTA. さらに フラグメント・セグメンテーション法を導入

2. 統計力学的: MDミュレーション (D.E.Shawの研究) や XTALポリス (モンテカルロ) 法

・ホモジー・モーリング法 (比較モーリング法)

オ1ステップ: フォールド認識 (BLAST 等または PSI-BLAST, HMMer 等のホール法を用いる) アライメント

オ2ステップ: モーリング テンプレートに基づいて全原子を配置 (Modeller, I-TASSER など)

⇒ 結果は詳細な原子構造を再現できないが 3D/4D単位のモーリングには有用、計算コストが低い

立体構造の分かっているもの: 3割り ホモジー・モーリング可能: 4割り 何もわからん: 3割り

3.5 分子間相互作用の解析

- ・ホモジー・モデリングみたいに... HOMCOS. 低分子-タンパクなら fkcombu など
- ・低分子-タンパクなら **分子ドッキング法**: (剛体)タンパクと低分子をポテンシャル的にドッキング DOCK, AutoDock Vina, ... オペ-キヤルスクリーニングも!
- ・高分子-タンパク: 形状(凸凹)や電荷(+-)の相補性を使つて MD など
並進(FFTで高速探索?)と回転でドッキングも. ZDOCK, MEGADOCK

§4. ncRNA 解析

4.1 ncRNAについて

- ・RNAの2次構造は抽象化した立体構造. 塩基対: A=U, G=C, G-T Watson-Crick Wobble

・短鎖ncRNA

miRNA (2600ほど by miRBase²⁰¹⁸) 複数のUTRにくっつけの2'. 6割のmRNAがmiRNAで抑制される?
piRNA (PIWI-interacting RNA) トランスポンタ由来 生殖細胞特異的(2種10万種↑) エピゲノムも関連?

- ・lncRNA: mRNAより多い説もある(16k) がん・神経疾患に関与か.



lncRNAの機能ドメイン: 構造モチーフ, リピート, RNA修飾, 他生体分子(タンパク, rRNA)との相互作用
 m^6A (N⁶-アザギニン), m⁵C, I, E, m'A, hm'c ...

4.2 ncRNAと大規模Seq解析

転写開始点の決定

- ・RNA-Seq, CAGE-Seq (Cap Analysis of Gene Expression)

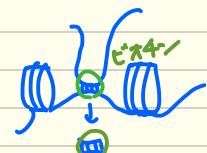
・相互作用解析

CLIP-Seq \Rightarrow RNA結合タンパクとの相互作用を1塩基解像度で ピピロストリック

近接ライティング + Seq \Rightarrow RNA-RNA相互作用



ChIRP (Chromatin Isolation by RNA purification) \Rightarrow RNAとゲノムrRNA相互作用



Ribo-Seq \Rightarrow lncRNAなどの転写産物のリボソーム結合領域の網羅的同定

- ・構造解析: DMS-Seqによる非塩基対性A,Cの同定などを用いて構造予測

・エピトランスクリプトーム: m'AなどのRNA修飾と機能の関係を研究

4.3 ncRNA解析のためのバイオインフォメティクス

浅井先生のアレ

- ・RNA 2次構造予測: MFE構造を確率分派自由エネルギー... より高いR \Rightarrow RNA composer, 共通2次構造予測 \Rightarrow RNA alifold ただし短いRNAだけ... 計算量エグい

- ・構造モチーフ \Rightarrow CM Finder (共分散モデル, やさしい)

- ・相互作用予測: CLIP-Seqを補うために. MLとか CatRapidなど

- ・既知ncRNAのゲノム検索: Infernal で

あとRNA-Seq, GSEA

GO enrichment von KEGGパスウェイも...

4.4 ncRNA 研究 (DBと例)

• ncRNA の DB

Rfam	500 bp 以上 ディファクトスタンダード
miRBase	miRNA ≥ pre-miRNA, 成熟 miRNA
piRNAdb	piRNA
GENCODE	ET/2クス ncRNA
Mitranscriptome) ET (ncRNA
Big transcriptome	CAGE-Seq に基づく lncRNA
FANTOM CAT	病気/SNPとの関連
NONCODE v4	色々な生きたもの
Expression Atlas	lncRNA 属性
Lnc ATLAS	lncRNA-RNA 相互作用
Lnc RR Idb	

• 例：抗がん剤薬剤耐性と lncRNA 解析

DEG → mRNA との co-expression & GO, KEGG → Survival analysis
生存時間解析

他にも
lncRNA 細胞特異的発現
とトランスポンの網羅解析
⇒ Fisher の正確確率検定

§5. NGS データ概論 4/28

5.1 NGSについて

• NGS データ

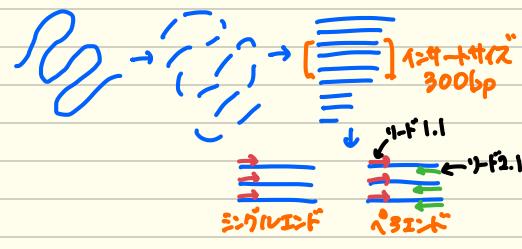
イニサートサイズ (フラグメントサイズ) : 解析断片長 一定の分布をとる 300 ~ 400 bp 程度

リード ⇒ 末端のみ読まれた配列) シングルエンド / ペアエンド

ロングリード (PacBio や Nanopore) ⇌ ショートリード (Illumina)
↑ リピート配列の解析などに強い ↑ リード数が多くなる

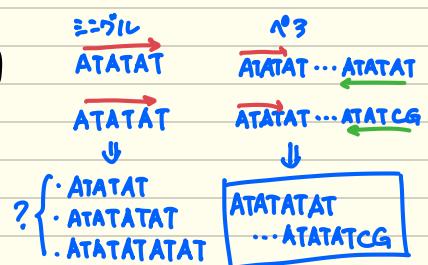
• 例 1: デノボラセンブリ (リードデータの組み立て)

⇒ コンティグ (組み上がり長い塩基配列) が得られる
たどりのリピート配列領域がクソややこしい ⇒ 分断しておく (シングルエンド)
ペアエンドではあいまいさなくリピート部分を推定! ? コンティグのかわりに
任意の塩基 N を含むことを...スキヤフォルドと呼ぶ



• 例 2: リニアクエンス

リフレンス配列と異なる箇所を変異として同定 (変異解析)
変異箇所の数 / NGS のエラー率の割合 = リード数



5.2 NGS データについて

• カバレッジ: (仮想)ゲノムサイズに対して使った全リードのデータの比率 ex. $\frac{72}{48} = 1.5 \times$

• クオリティスコア: ベースコール (NGS の塩基決定) におけるエラーの指標

$$Q = -10 \times \log_{10} P \quad \Rightarrow Q_C \text{ (クオリティコントロール) すべし} \leftarrow FastQC や trim galore! で見て
(NGS は読みほどクオリティが下がっていく)$$

• 形式: SRA / FASTQ

4テレ 1リード分

72 bp × 122
48 bp × 122

§6. ゲノム解析

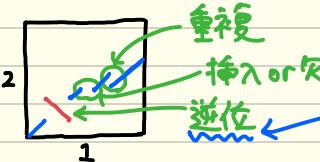
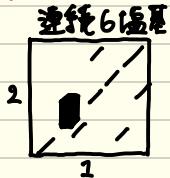
6.1 ゲノム解析で分かること

$$\text{GC content} = (G+C)/(A+T+G+C)$$

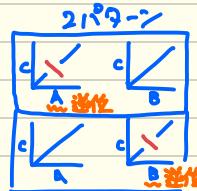
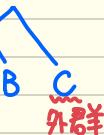
$$\text{GC skew} = (G-C)/(G+C) : 片側の鎖のGC偏りの指標, Sliding Window を用いる。$$

かつては「移動平均」のこと

- ドットプロット：2つのゲノムを比較



どちらのゲノムが挿入・欠失も!
逆位を起こした?
⇒系統樹的に判断!



重複はキボン見れば分かるけど、不等交叉も考えるなら、外群便った比較で

ヒトと2ウツの dot plot にはシンテニー（オーログの並びが類似していた）領域が多く見られる

- 比較ゲノム解析：ズルテアルスラインメントによる

- 配列ロゴ表記：各サイトの塩基の表れやすさを情報量で示す。

$$C_i = 2 + \frac{\sum p_{a,i} \log_2 p_{a,i}}{\text{max値}} \quad p_{a,i} : \text{サイト } i \text{ の塩基 } a \text{ の割合}$$



6.2 ゲノムブラウザ

- ゲノムブラウザ：アノテーション（附加情報、どこにSNPなどの差の出る場所があるかなど）を含むDB

UCSCゲノムブラウザやEnsembl

⇒ "Genomes" でセンブリバージョンを選べ、検索とかをする

<ヒトゲノム>

UCSC NCB

hg38 = GRCh38

> 19 = : 37

: :

§7. トランスク립トーム解析

7.1 RNA-SeqとQC

転写物

- 発現解析：スライスペリントの発現量の違い・変化を見る

- アダプター配列：(サイズの決まり) cDNA 末端に付加、PCRのプライマー用とか ⇒ リードに含まれる!?

⇒ 非ゲノムな配列をQCで除去 → 3'アダプター配列をラインメント

(23) ATTCTAGCATAG
+5 AGGATAGCCTACA
-7 AGGATAGCCTACA

7.2 マッピング

- ショートリードのゲノム配列へのマッピング

① 1つのエクソン内に完全マッチ

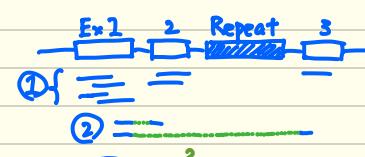
② 2つのエクソンにまたがるジャンクションリード

リピート配列マッピング(ちゅうごんも... ⇒ リピートをNでマッピング!)

- 数値化

マッチ数

⇒ 共有エクソンの判定が難しい



←結果から
新規転写物を
同定することも
75bp分もあるから!



発現変動の群間比較は多重比較問題

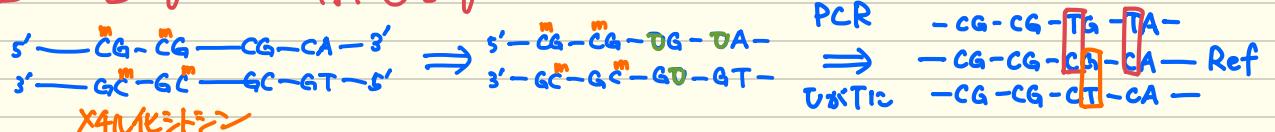
§8. エピゲノム解析

4/29

- 正常エピゲノムのデータ ⇒ THEC Data Portal
- 異常エピゲノムのデータ ⇒ ICGC Data Portal

8.1 エピゲノム解析の計算手法概説

• BS-Seq (bisulfite sequencing)



• ChIP-Seq

免疫沈降により選択され集中してピーカーを検出する。
MACS2, Homer, MAnorm, ChIPComp, DiffBind

生物統計論の復習しよう!!
MACSの仕組み

8.2 解析の実際

• BS-Seq

bisulfite処理のために一定量の試料があることが肝要。業者・機関に託すのは委託。

- モルヒ生データ ⇒ QC FASTQC + MultiQC ↗ Cが下がってTが上がり
- 3'ドロップ-配列除去 Trimmomatic / Cutadapt (Trim-galore!)
- 2-ヒビング moLPCR EL2などの影響が出るなら (BISMARCK) deduplicate_bismark
- X4U変換率推定 bismark-methylation-extractor
- 複数試料で比較し、DMR (X4U化の変化領域) を抽出 ComMet, bsseq
- 関連遺伝子群の抽出やGO enrichmentなど... "GREAT"が使えるWebツール

• ChIP-Seq (1~3章一緒)

- ピーカー検出 MACS2など → DPR (ピーカー変化領域), DEG (発現変動遺伝子) を
- GO enrichmentなど... DAVIDなどが使える

§9. メタゲノム解析

5/1

9.1 メタゲノム解析とその種類

- 環境中の微生物群集から抽出したDNA全体をシーケンスして解析

ヒト共生微生物群集 → HMP, MetaHIT

- 解析手法 2つ

アンプリコン解析 ← 基礎論Ⅱのラストで扱ったやつ

16S rRNAなどをPCRで増幅した産物(アンプリコン)を系統解析

↑ 保存領域と可変領域が同定されてる。しかし、300bpほどのリードでは属するしか分からない

PCRのプライマーはユニバーサルプライマー(可変領域を含むように)

1. プライマーの2倍鎖となる領域をペアエンドにシーケンスして合体する paired-end reads merge
→ 長くなる!

2. QC (3'フラット・プライマー・キメラ配列の除去)

PCRエラー UCHIME2, ChimerSlayer

3. 配列クラスタリング ⇒ OTUの作成 ← 100%相同な配列を作ることが多いらしい

4. 配列相同性検索・系統学的解析 BLASTより高速な RDP, SILVA, Greengenes

ショットガン解析 (ぶちがちにて Seq.)

1. メタゲノムサンプル 普通のサンプルに対する問題点を加味する

2. 遺伝子予測 [ラビングで得られたコテイグ] 断片配列から) あらゆる 不完全で多様 GCやk-mer頻度を利用?

3. 既知ゲノム配列への匹配スッピング

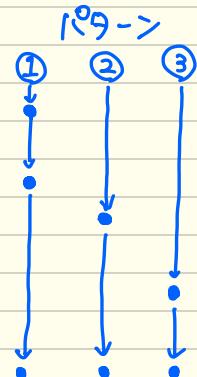
4. 遺伝子機能予測 GenBank, RefSeq, KEGG ...

- ゲノム再構成の手法

Binning : 細菌種ごとに Bin に分類し。Binごとに構築

↳ コテイグの k-mer パターン, コンティグカバレッジの変動から!

サンプル → k-mer / カバレッジ → EM モル比で分類 → Bin のエンコード, キメラなどのクリーニング



9.2 比較メタゲノム解析

- 複数サンプルの比較

最も少ないものの数だけサンプリング

1. 变異の正规化 : 分類された配列を割合に ⇒ rarefying

2. サンプル間距離行列 : Bray-Curtis 非類似度など... ⇒ 補正したのが UniFrac

3. 距離行列可視化 : PCA or MDS (多次元尺度構成法) を利用

§10. プロテオーム解析

10.1 プロテオームとは何か

- タンパク質の発現状況は遺伝子の Reference 配列だけでは分からぬ
塩基配列の揺らぎ(個人差), 突然変異, PTM (翻訳後修飾) ...

⇒ プロテオーム (isoform, protein variant, PTM を含む バリエーション全体)
proteoform
の全体をプロテオームと呼ぶ

- プロテオームのバイオインフォメティクス

プロテオフォームは (PCRのように) 増幅が出来ない! ⇒ 生物物理的な解析

[計算プロテオミクス] : プロテオミクスのための計算・解析手法の研究

[プロテオーム情報学] : PPI (タンパク質間相互作用) に代表される研究

10.2 インタラクトーム解析 {PPI}

- インタラクトーム (相互作用ネットワーク) の実験解析

A. 体構成系

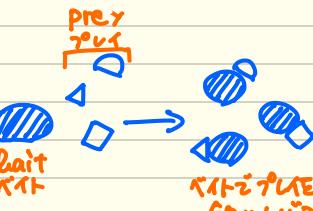
酵母ツーハイブリッド法 Y2H

プロテインアレイ

フローシーティングアレイ

遺伝子組換え

A-α
B-β



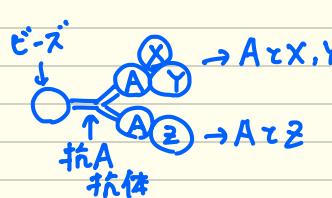
Bait
Prey

表面プラスモニ共鳴
発光などを検出

B. in situ 系

Co-IP

エピトープ・タグ
アッダーウン



→ 全部の方法で 質量分析へ
MS

3つ以上の
会合 PPA
も判定できる!!

共分画 (MS + きかいがくじゅう)

- インタラクトーム理論解析

論文のアブストから PubMed 上での テキストマイニング

- オペロン構造の保存性からの推測
⇒ 真核生物では 系統プロファイル から!

- ロゼットストーン法 <遺伝子が融合してみた>
可探可

- データベース

STRING, HPRD, IntAct, ...

系統プロファイル				
A	B	C	D	E
1	2	1	1	0
0				
2	2	0	2	1
1	2	0	2	1
1	1	1	1	1

弱点
・決定にゲル
・オーソログ判定に依存
・全ゲルムにあるものにのみ

内積が大きいと
相互作用している?

10.3 プロテオームの同定

2次元電気泳動もあれば

・質量分析による測定 ショットガン・プロテオミクス

1. 前処理 : affinity chromatography や抗体などでタンパク質を分離

2. 質量分析計でイオン化 (ESI や MALDI) ・ 分析計 ・ 検出器を通って計測

m/z 値が手に入る

MS/MS

3. 実際は以下のように「解離反応室」を狭んでもう1つ分析計を入れる。

ペプチド切断

イオン源 → 分析計 → 解離反応室 → 分析計 → 検出器
MS precursor ion production MS/MS ← 部分構造の測定!

計算アロテオミクス

4. ペプチドの波形処理とピーコ検出
横軸 m/z 縦軸 依存流強度 スペクトルは山型
ある m/z をもつペプチドの時間変化
⇒クロマトグラム

5. ペプチド同定 スペクトルごとのイオン電流とペプチド量に比例するとは限らない!!

スペクトルライピング法や de novo sequencing 法などがあるけど、
データベース検索法 PMF が主流(らしい) あるいは MS/MS ion search も!

6. タンパク質推定 UniProtなど... ただし、ペプチド → 3ペプチドが 1対1では限らない

7. 信頼性評価 : 多重検定と FDR ← Target-decoy 検索

(8. タンパク質の定量 : emPAI といったカウントデータへ強度を対応、半定量的)

§ 11. データベース

各種データベース自体の話は Web にまかねば

○ FAIR : Findable, Accessible, Interoperable, Reusable
相互運用性

○ ファイル形式

CSV かんたん
TSV タブ区切り

FASTA

フラットファイル

← PDB や GenBank などでの
TogowS など
独自の形式の総称

○ データベースシステム

RDB (Relational Database) : 表形式のデータの関係で扱う。管理システム DBMS の内閣せ言語は SQL
MySQL, PostgreSQL, SQLite ... Ruby を使うことも

NoSQL : SQL を使わない DB. Key-value ストア, MongoDB, Solr, Elasticsearch など...

RDF (Resource Description Framework) : Semantic Web のデータモデル。トリプルで記述
SPARQL 言語で問い合わせる。NoSQL で唯一標準化

○ Web API

バイオインフォメティクス : NCBI E-utilities, ENSEMBL REST API, TogowS ...

§ 12. バイオのためのML

12.1 確率モデル

- 生成モデル：観測変数を隠れ変数で表現

HMM：エピゲノムデータでクロマチン状態を推定など
浅井先生から習ったテレ
Viterbi や Baum-Welch など

PHMM：ペア隠れマルコフモデル、ペアワイズアラインメントを利用。シーケンスなどの特徴を学習したりでき
PDTA-HMM "モデル学習"

SCFG 確率文脈自由文法：RNA 2次構造などを $S \rightarrow aSu$ のように表現

共分散モデル：SCFG 版の PDTA-HMM ↑

トピックモデル PLSA 確率的潜在意味解析 ← 気になる!
LDA 潜在ディリクレ乱択モデル
例：タケミツ細菌の解析・がんの変異シグネチャー "モデル学習"

- その他手法

SVM・カーネル法

ランダムフォレスト

スリットモデル 深層学習

CNN・RNN・LSTM

理屈とかの授業を踏まえて

じっくり学ぼう!