資料3－4

2008年1月8日

ライフサイエンス統合データベースプロジェクト
共通基盤技術開発チーム進捗状況報告（CBRC）

CBRC：浅井、野口、諏訪、
金、ホートン、広川、福井、藤渕、堀本

第3回統合データベースプロジェクト
研究運営委員会資料

担当業務：ライフサイエンス統合データベース開発運用 (共通基盤技術開発)

ワークフロー技術を用いた統合DB環境の構築

利用者が得たい情報（知識）を、パソコンなどの端末から要求すると、必要なデータ、解析手法などを、国内、海外から自動的に選び、データベースと解析ツールのワークフローを作成し、最適な計算資源を使って解析を行う統合DB環境を構築する。

統合データベースのイメージ

例　この薬の副作用を予測したい

化合物
＋
本来の標的遺伝子

最新のIT技術を利用し、必要なデータ、解析手法を国内、海外から自動的に選びパイプラインを構築、最適な計算資源を使って解析を実行して返す

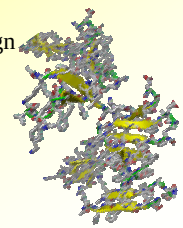
相互作用･結合DB検索
構造予測ベース解析
ターゲット遺伝子予測
細胞内ネットワーク解析

目的外標的遺伝子リスト
分子設計指針
副作用予測結果

バイオインフォマティクス
統合環境ウェブページ

Integrated Environment for Bioinformatics Knowledge Discovery

- Structure based drug design
- Target gene screening
- Side effect prediction



DB検索要求
データ解析要求
結果取得

DB検索要求
結果取得

国内サイトのDB

独自DB、ツールと
大規模計算を組み合わせた
解析要求･結果取得

データベース群

統合データベースセンター

Mirror Site: DDBJ, PIR, PDB, SWISS-PROT, KEGG, etc.

INOH　fRNA db　EzCatDB　JSNP　ProMode　SEVENS

ツール群

ASH　BLAST　CoCoozo　SCARNA　GRIFFIN　POODLE　WoLF PSORT　ALN　FORTE　CELL MONTAGE　ASIAN

計算資源

統合ＤＢにおけるワークフローの開発方針

- **グリッドを利用**

  理由１：目標とする統合ＤＢを構築するために必要な技術が、ほぼグリッド環境で実現できている。（例：ＧＥＯGrid）

  －長時間JOBの対応

  －ソフトウェアやデータベースの利用制限（アカデミック利用のみ等）

  －その他

  理由２：ワークフロー実現のために必要で、かつ現状のグリッド技術にない技術の開発に関しては、産総研グリッド研究センターの協力が得られる。

- **ＣＢＲＣ内でプロトタイプを構築し、外部と連携**

今年度の計画

- CBRC内の計算機環境で、固定のデータベースと解析ツールを用いたワークフローを作成し、統合DB環境のプロトタイプを構築する。

ーグリッド環境構築

ーワークフローの議論

ーソフトウェア・データベース間のインターフェイス

共通基盤技術開発チーム内の打ち合わせで、SOAPに決める

→ＣＢＲＣにおいて公開しているＷｅｂサービスのSOAP

①ワークフロー固定

②ワークフロー可変 (taveruna, KNIME調査)

ーグリッドWebアプリケーションの開発

システム構成

産総研外ソフトウェアおよびDB

標準的なWeb Serviceで統合

OGSAによる標準アーキテクチャ

グリッド環境の計算資源への計算依頼

データグリッドによる大規模データの仮想化 OGSA-WebDB

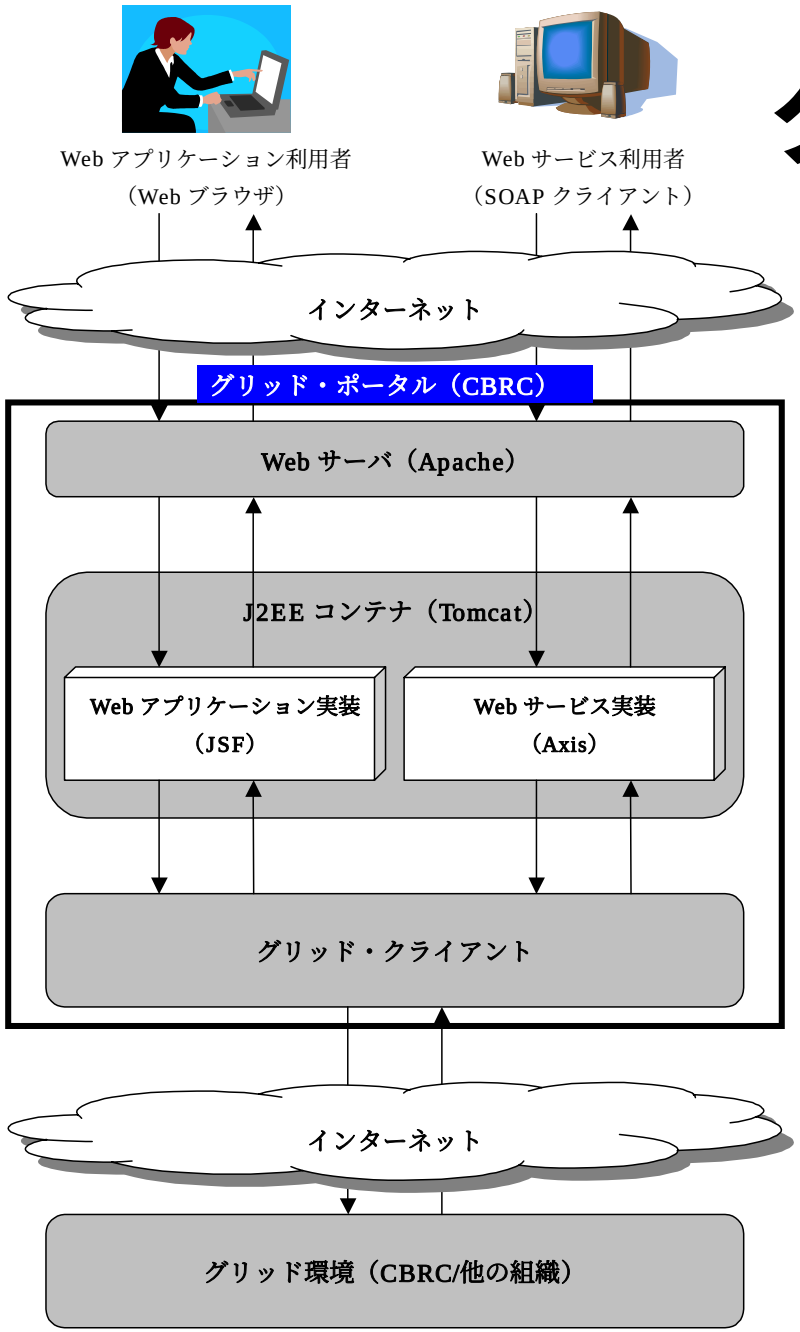
産総研内計算資源

グリッド用計算サーバ、Magiクラスターなど



産総研内DB

グリッド環境構築



ハードウェア：
- グリッド用Webサーバ
- グリッド管理サーバ
- 計算サーバ（発注済）

ソフトウェア：
- OS(CentOS 5.0)
- Web関連
- グリッド関連

ＣＢＲＣで研究開発された公開ソフトウェア（例）

Software

ALN
2つの配列または配列グループ間を
http://www.cbrc.jp/ALN/

ASIAN
階層クラスタリングとグラフィカル・ガウ
み合わせによるネットワーク推定ソー
http://eureka.cbrc.jp/asian/

CellMontage
マイクロアレイデータ検索・解析システ
http://cellmontage.cbrc.jp/

CoCoozo
ペプチド・タンデム質量分析向け並列
http://www.cbrc.jp/cocoozo/

FORTE
プロファイル比較によるタンパク質立
http://www.cbrc.jp/forte/

GeneDecoder
真核生物ゲノムのDNA塩基配列から遺伝子を予測するソフトウェア
http://genedecoder.cbrc.jp/

GRIFFIN
GPCR－Gタンパク質結合選択性予測シス
http://griffin.cbrc.jp/

GUPPY
遺伝子配列におけるデータの意味を註釈
グラム
http://www.cbrc.jp/GUPPY/

Murlet
構造RNAのマルチプルアライメントプログ
http://murlet.ncrna.org/

PAPIA
並列タンパク質情報解析システム
http://mbs.cbrc.jp/papia/

POODLE
タンパク質ディスオーダー予測
http://mbs.cbrc.jp/poodle/

Scarna
類似RNA高速検索ツール
http://www.scarna.org/

SGCAL
糖鎖断片化解析ツール
http://sgcal.cbrc.jp/

SOKOS/CAN
多目的RNA配列解析ツール
http://www.cbrc.jp/sokos/

TMBETA-NET
膜貫通タンパクのベータストランド予測をするプログラム
http://psfs.cbrc.jp/tmbeta-net/

WoLF PSORT
タンパク質細胞内局在化予測ソフト
http://wolfpsort.org

その他の主な公開ソフトウェア

- **RNAmine** http://www.scarna.org/rnamine/ （SOAP化済）

～RNA配列からの2次構造モチーフ探索サービスを提供している。複数のRNA配列（構造アノテーション不要、アラインメント不要）を入力すると、そこから2次構造モチーフを抽出する。独自のアルゴリズムによって、極めて高速なモチーフ抽出を実現 しているところが特徴。

- **Murlet** http://murlet.ncrna.org/ （SOAP化済）

～RNA配列用高速多重配列アラインメントサービスを提供している。複数本のRNA配列（構造アノテーション不要）を入力すると 2次構造を考慮した配列アラインメントの結果を出力する。 独自の工夫によって、従来の手法に比べて高速な処理を実現しているところが特徴。

- **PHMMTS** http://www.scarna.org/phmmts/ （SOAP化済）

～RNA配列用構造アラインメントサービスを提供している。構造アノテーション付RNA配列（複数本可）と、構造未知のRNA配列1本を入力として、構造未知のRNA配列に対する構造アノテーションを推定する。RNA2次構造情報を木構造と隠れマルコフモデルの組合せで表現するという新しいアイデアを実証するための proof of concept 的サービス。

- **MXSCARNA** http://mxscarna.ncrna.org/

～RNA配列用超高速多重配列アラインメントサービスを提供している。 複数本のRNA配列（構造アノテーション不要）を入力すると 2次構造を考慮した配列アラインメントの結果を出力する。独自のアルゴリズムによって、従来の手法に比べて極めて高速な処理を実現し、かつ精度も同程度を維持しているところが特徴。

ＣＢＲＣで研究開発された公開データベース (例)

Databases

ASTRA
選択的スプライシング・選択的転写開始パターン分類データベース
http://alterna.cbrc.jp/

ConfC
タンパク質構造変化部位データベース
http://mbs.cbrc.jp/ConfC/

DB-SPIRE
タンパク質機能構造データベース
http://mbs.cbrc.jp/DB-SPIRE/

EzCatDB
酵素触媒機構データベース
http://mbs.cbrc.jp/EzCatDB/

fRNA Database
fRNAdbとUCSC GenomeBrowser for Functional RNAで構成される新規機能性RNA遺伝子発見支援用データベース
http://www.ncrna.org/

GENIUS II
全ゲノムタンパク質の立体構造帰属データベース
http://genius.cbrc.jp/

INOH
シグナル伝達パスウェイ・代謝パスウェイデータベース
http://www.inoh.org/

PDB-REPRDB
配列、および構造類似性を考慮した代表タンパク質チェインデータセットを作成するシステム
http://mbs.cbrc.jp/pdbreprdb/

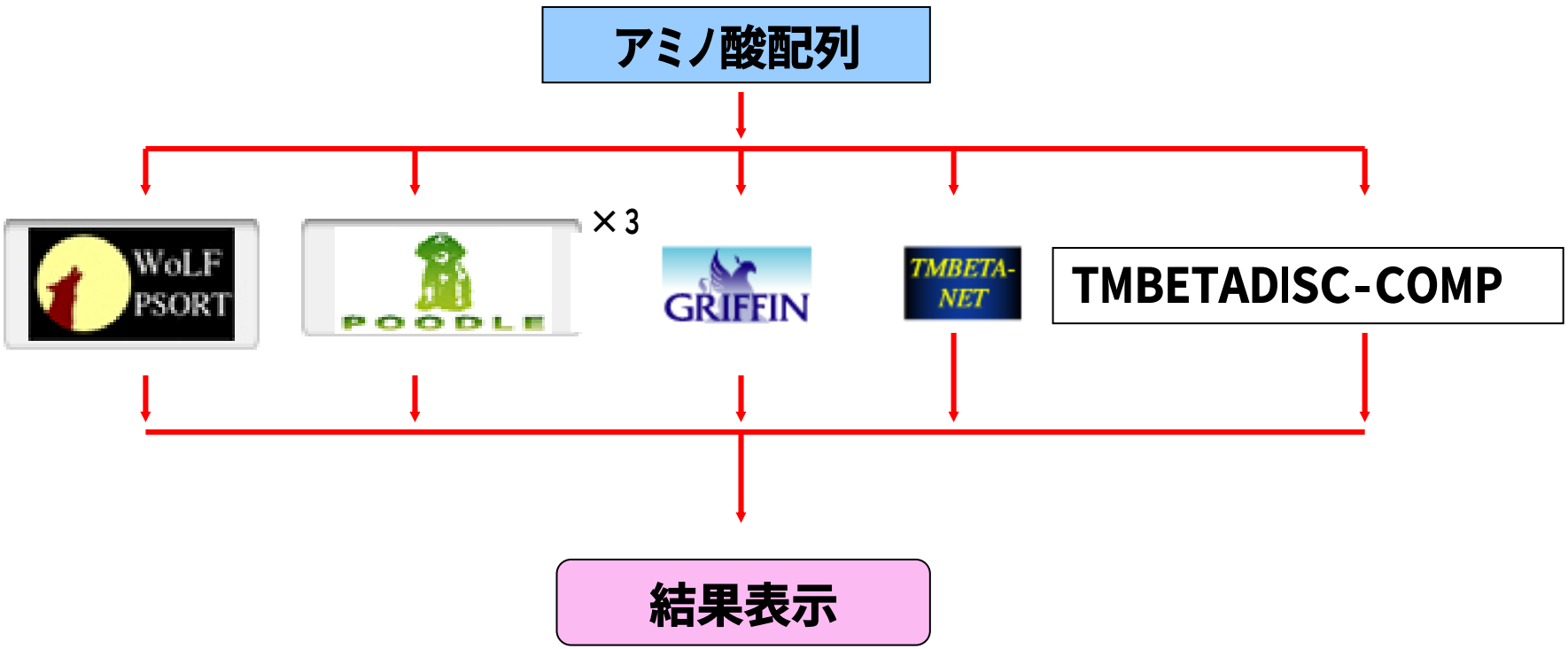
SEVENS
GPCR遺伝子の網羅的データベース
http://sevens.cbrc.jp/

TMBETA-GENOME
ゲノム中に含まれるβ－バレル型膜タンパク質のデータベース
http://tmbeta-genome.cbrc.jp/annotation/

CBRC内公開サービスのSOAP化状況

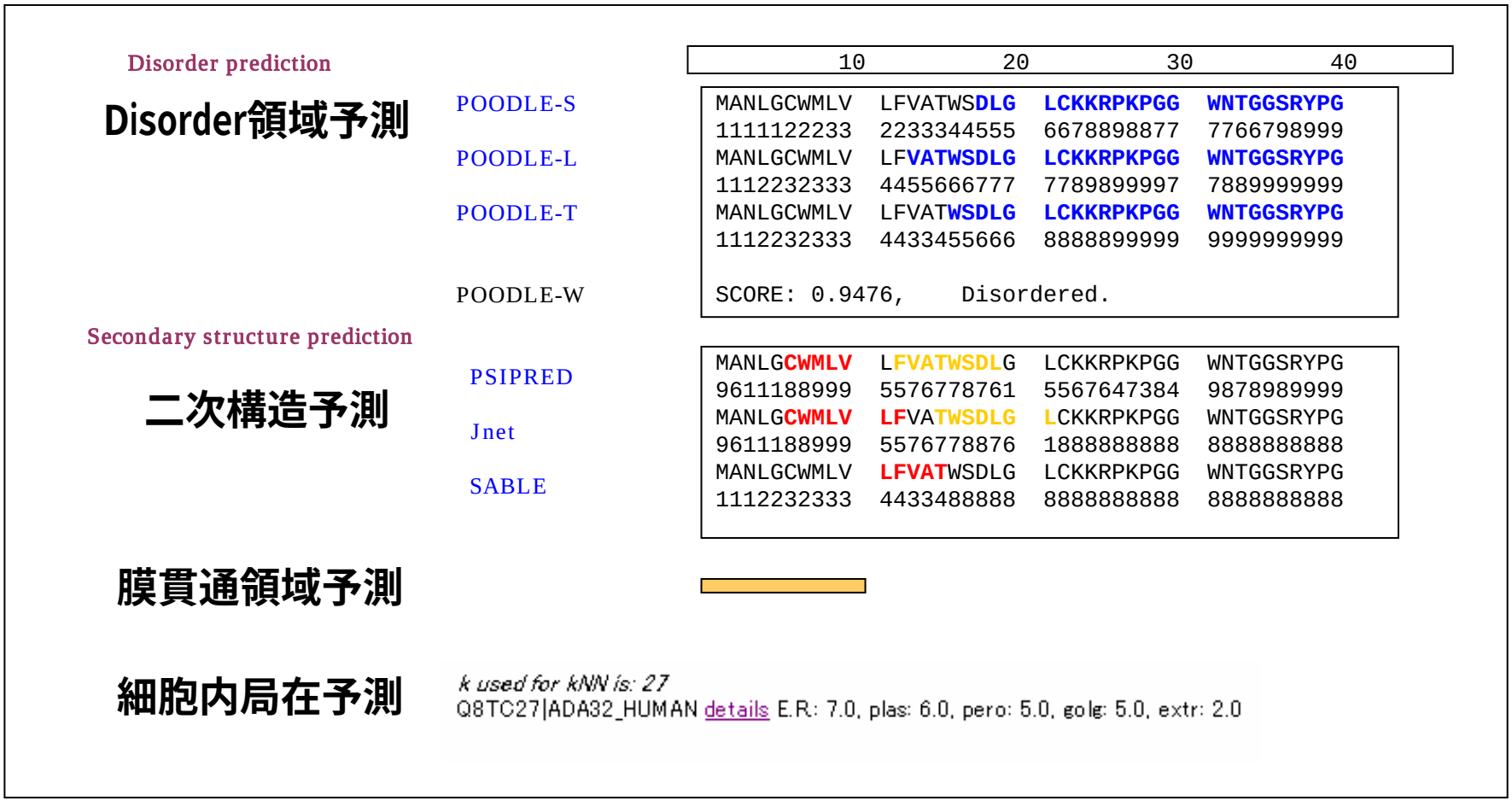
	計
SOAP化されたDB	0
SOAP化されたソフトウェア	4
SOAP化中のDB	0
SOAP化中のソフトウェア	7
SOAP化未対応のDB	10
SOAP化未対応のソフトウェア	15

プロトタイプ1 (配列解析)



注) DBCLSの予算でSOAP化

プロトタイプ結果表示イメージ

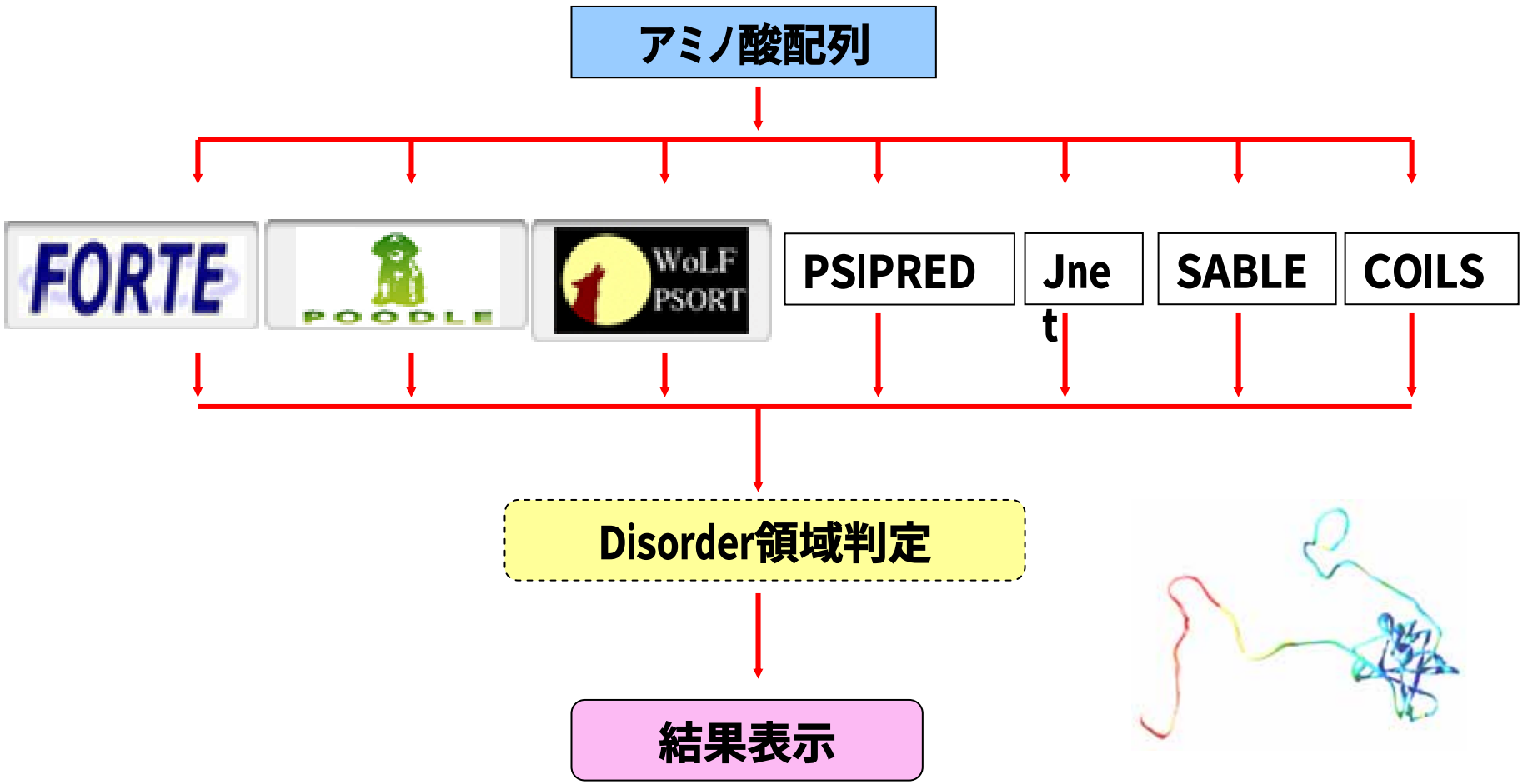


CBRCの公開Webサーバーの結果をこのイメージで貼り付けていく

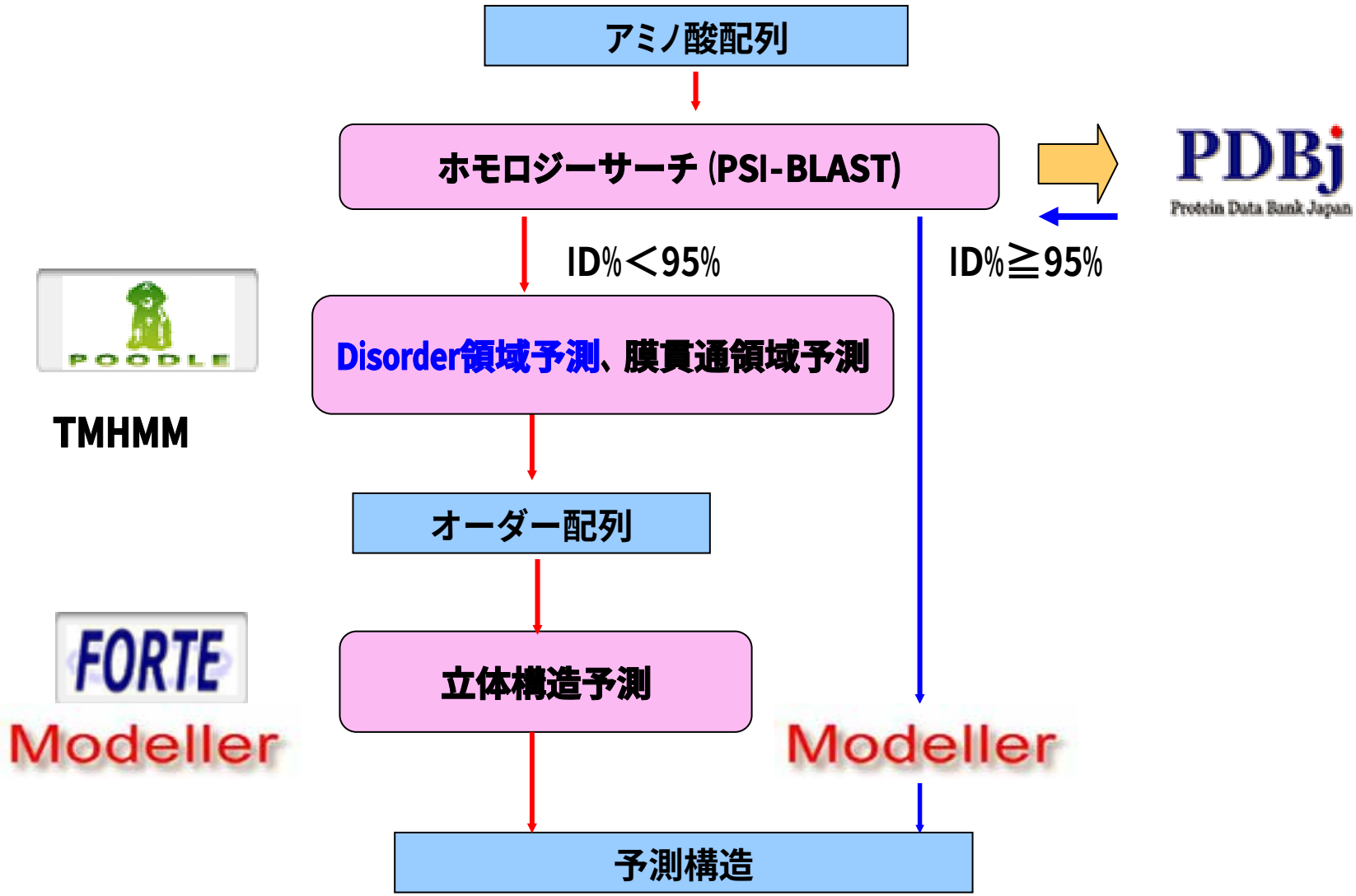
プロトタイプ2

（CASP7で用いたDisorder領域予測手法）

～総合評価で2位、長いDisorder領域で1位～

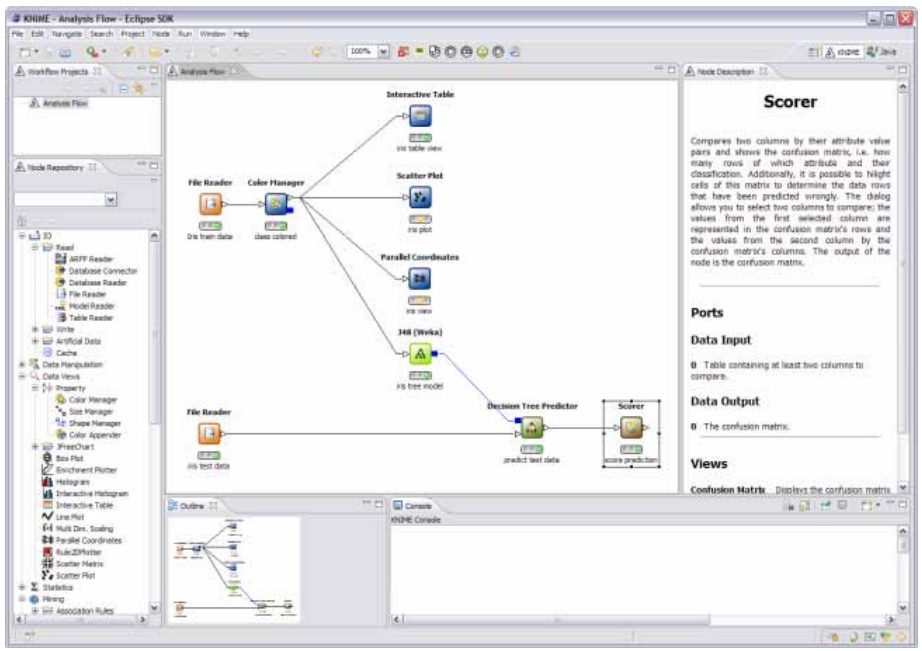


タンパク質立体構造予測ワークフロー

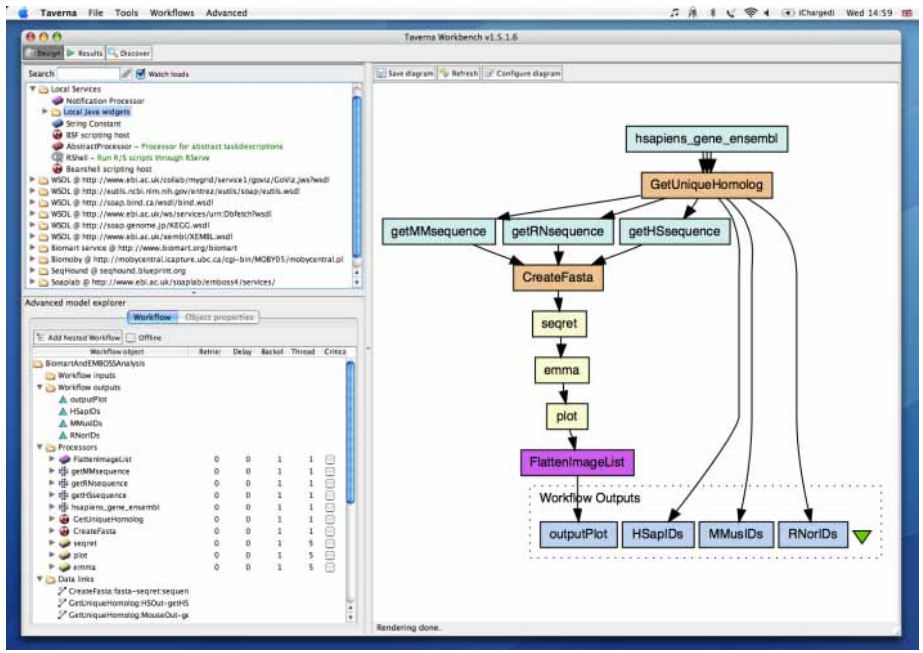


注) JST BIRD「タンパク質の構造･機能予測法の開発とヒトゲノム配列への適用」(代表：東工大　太田元規准教授)
において研究開発中

ワークフロー技術調査



KNIME

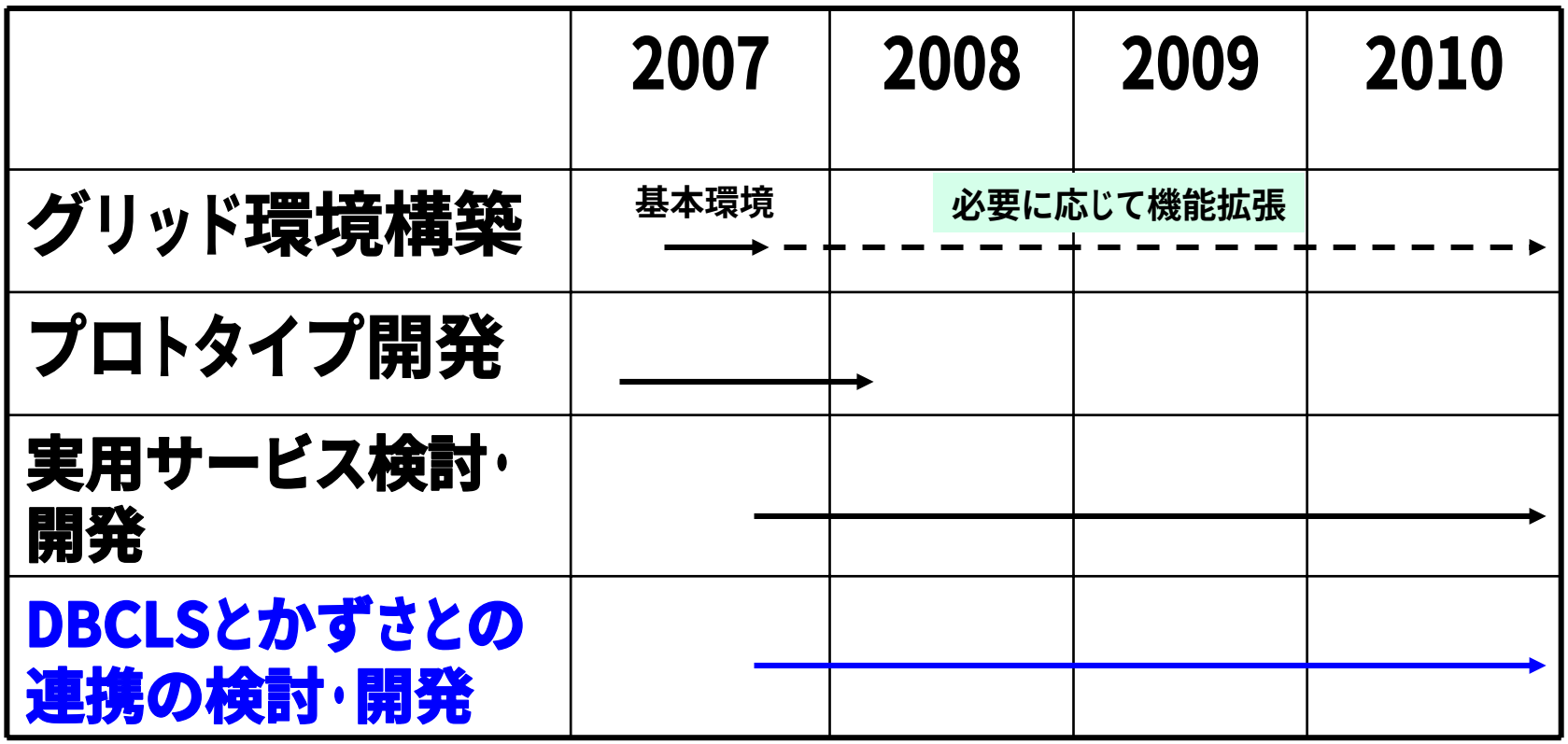


taveruba

今後の予定

- **DBCLSとの連携**
  1) CBRC のサービスが DBCLS のサーバから利用可能
  2) CBRC のサービスにDBCLSのサーバのサービスを組み込む
  3) 上記以外の連携は、定期的な打ち合わせを行い議論
- **CBRC内のグリッド環境**
  1) DBCLSとグリッド環境の連携
  2) グリッド用計算サーバ整備
- **ワークフローのサービス**
  1) 他のサービスの検討･設計
  2) ワークフローを構成するソフトウェアのモジュール化
     (taveruna、KNIMEのイメージ)

開発計画


2007
2008
2009
2010
グリッド環境構築
基本環境
必要に応じて機能拡張
プロトタイプ開発
実用サービス検討・開発
DBCLSとかずさとの連携の検討・開発