資料 5

2009年5月25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)

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)

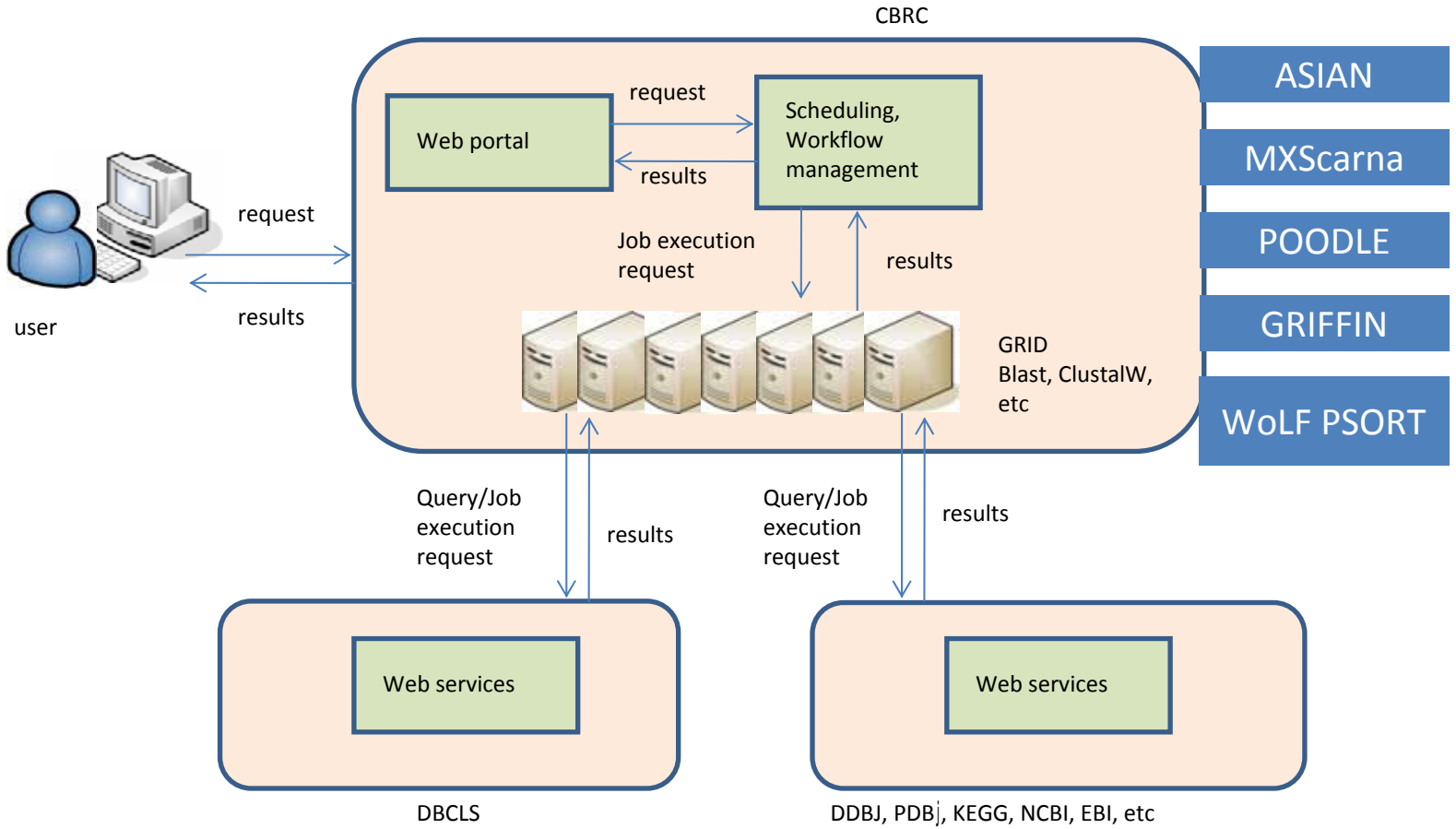
ワークフロー技術を用いた統合DB環境の構築

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。

CBRC：浅井、野口、諏訪、
福井、光山、ホートン、広川、藤渕、堀本、田代

2007年度～

CBRCは、２００７年度よりCBRCの各サービス（データベース及びソフトウエア）と他の有用公共サービスを統合したワークフロー構築とそのワークフローの効率的な分散処理環境の整備に注力。



2008年度実績

２００８年度は、２００７年度に実施した調査･環境構築･プロトタイプ構築を基に3つのワークフローを開発

⑴ タンパク質構造情報ワークフロー
（Protein Structure Information Workflow）

⑵ タンパク質アノテーションワークフロー
（Protein Annotation Workflow）

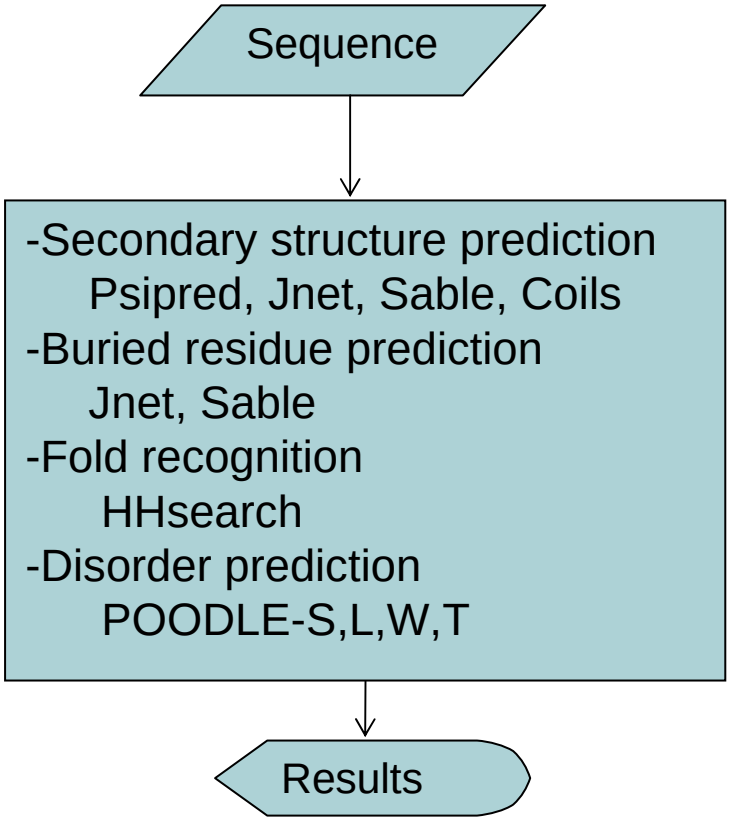
⑶ タンパク質比較情報ワークフロー
（Protein Comparative Information Workflow）

(1) タンパク質構造情報ワークフロー
(Protein Structure Information Workflow)

- タンパク質立体構造に関する予測プログラム等をGridにより効率的に分散処理し、非分散時の1/5以下の処理時間で行う。

- ユーザとしては、立体構造未知のタンパク質に関し、何らかの構造上のヒントとなる情報を必要とする実験研究者等を想定。

- 2008年8月29日に、プロジェクト関係者に限定し公開。 次バージョンワークフローの一部として位置付け。



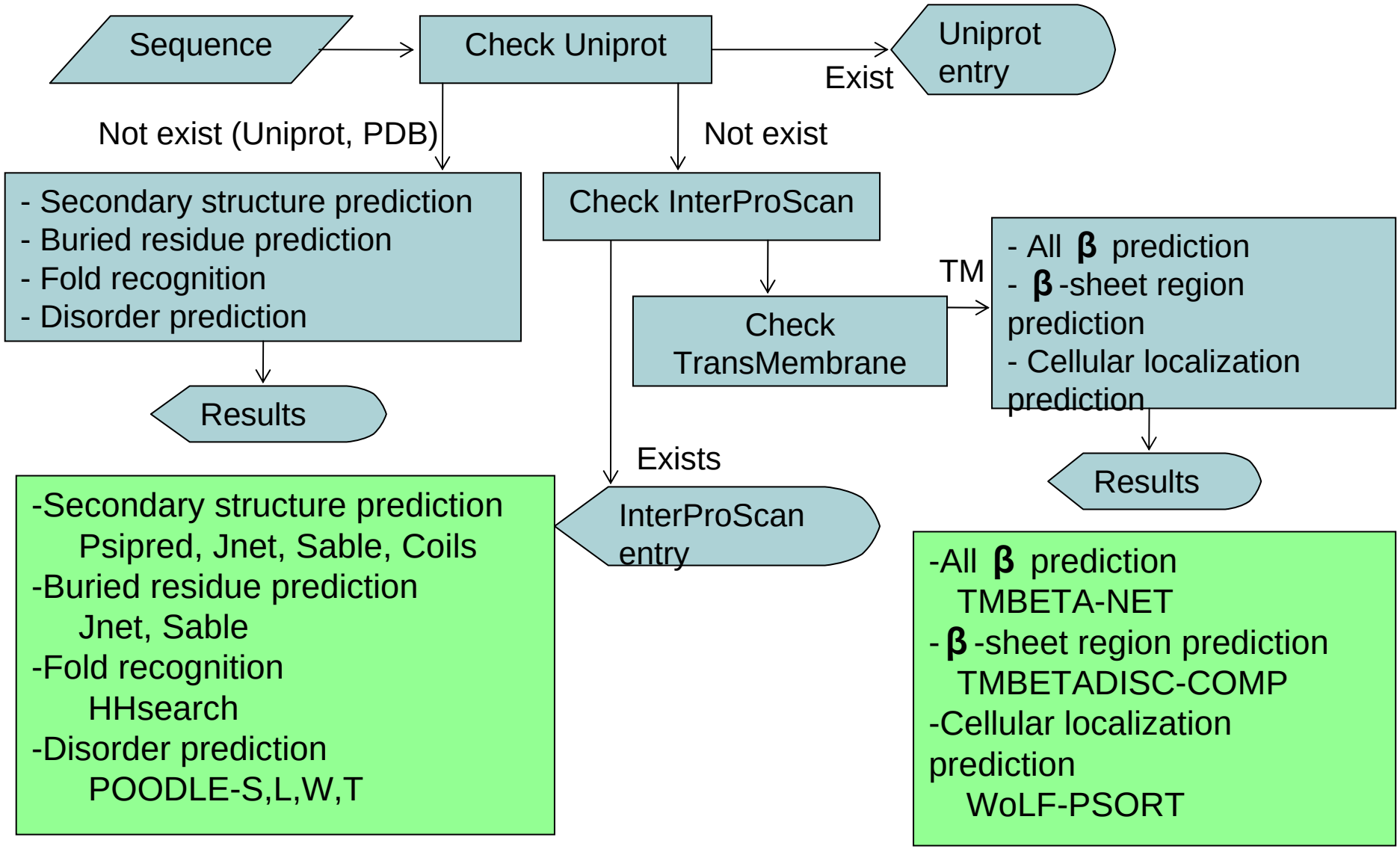
(2) タンパク質アノテーションワークフロー
(Protein Annotation Workflow)

- 想定ユーザは (1) のタンパク質構造情報ワークフローと同様であり、(1) を発展・拡張。(1) と同様に各種プログラム等をGridにより効率的に分散処理し、従来と比し短時間で結果を表示。

- ユーザからアミノ酸配列を受取り、二次構造予測、埋れ残基予測、フォールド認識、ディスオーダー予測、膜タンパク質オールベータ・ベータシート予測、細胞内局在予測を分散処理する一方、データベース検索及び疎水性予測の実行を他のサーバへ依頼し結果を取得後、全ての結果をユーザが解析し易いよう配置し出力。

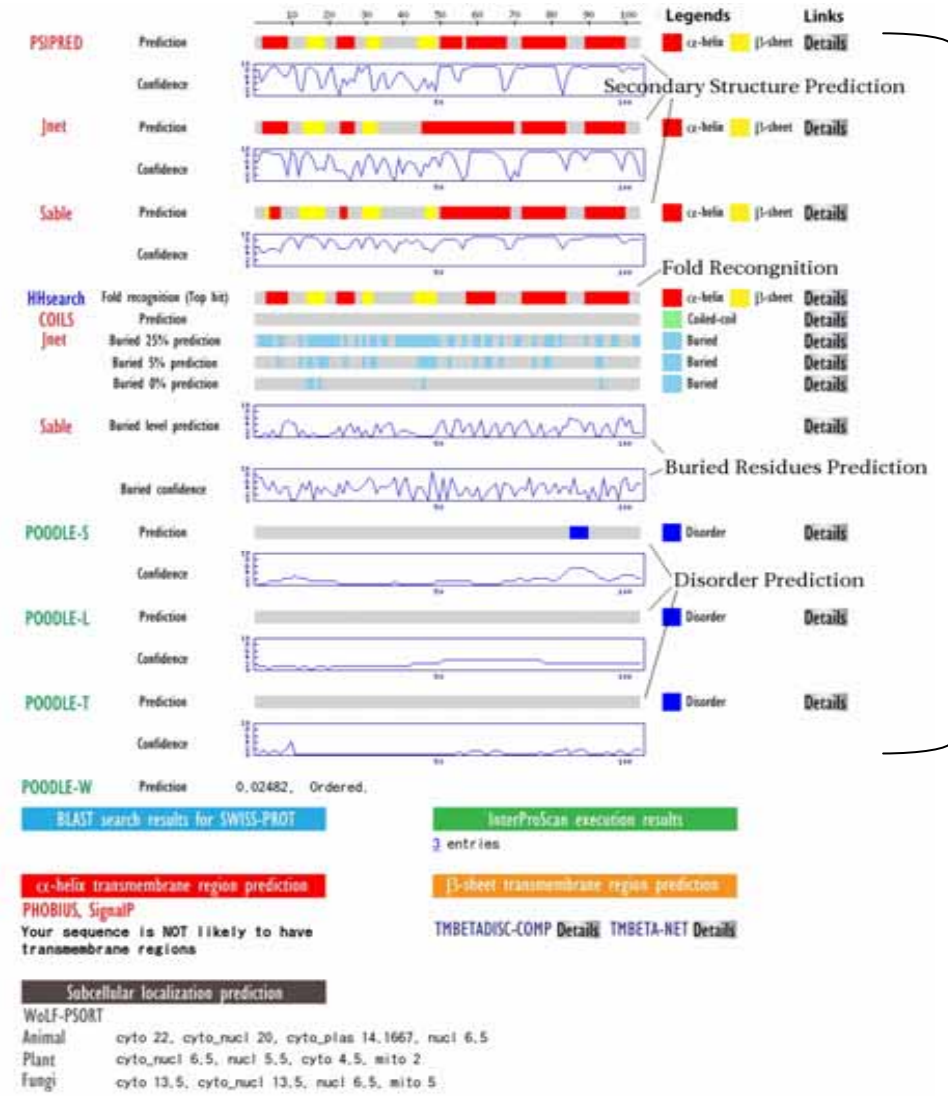
- 2008年12月27日より一般公開。

(2) タンパク質アノテーションワークフロー
(Protein Annotation Workflow)



(2) タンパク質アノテーションワークフロー
(Protein　Annotation　Workflow)

結果画面



異なるプログラムの予測結果を揃えて表示し、目視認識度を向上

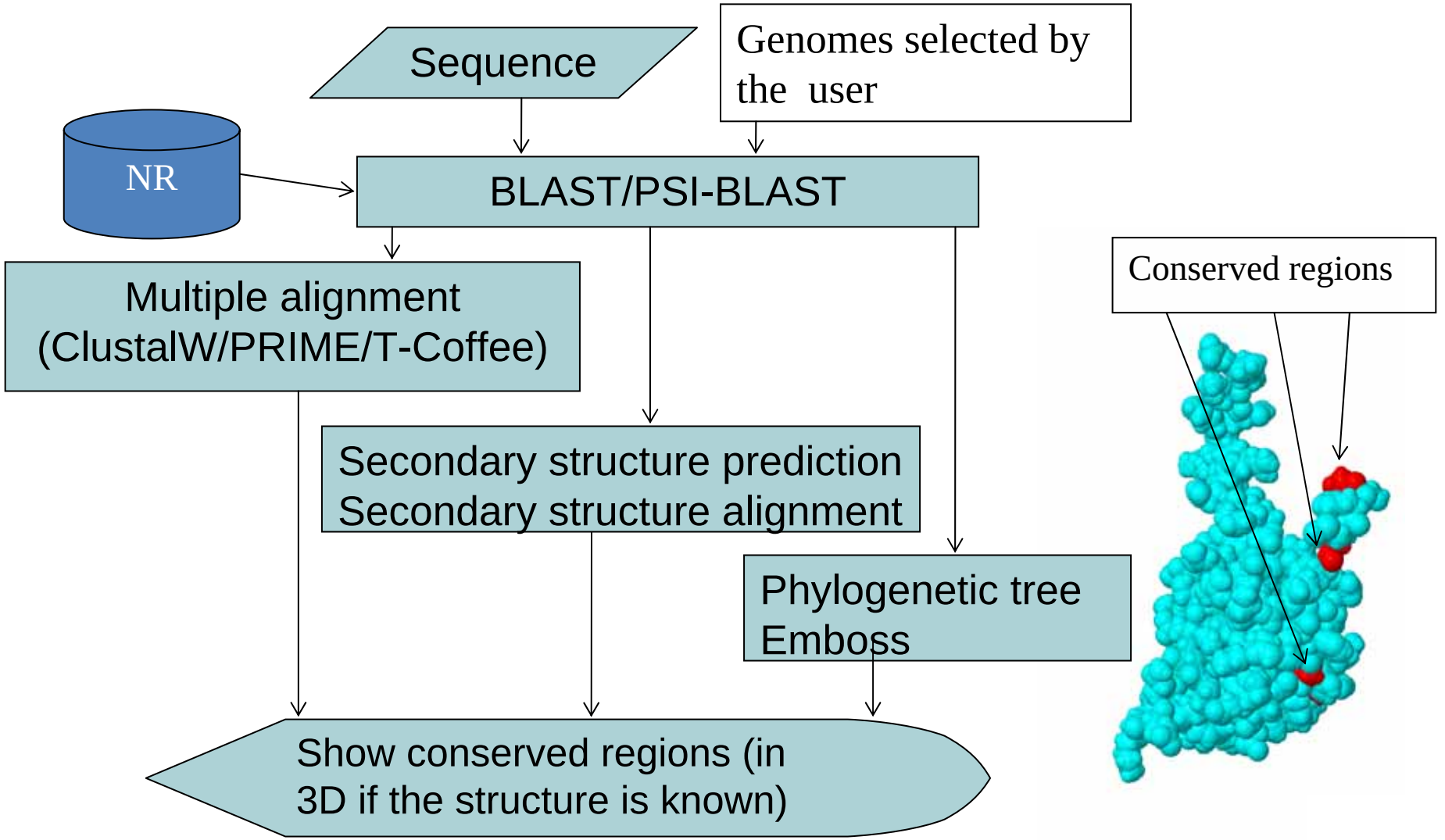
(3) タンパク質比較情報ワークフロー
(Protein Comparative Information Workflow)

- 相同なタンパク質を比較することで保存部位等構造上重要な部位を表示し、実験研究者等に提供することが目的。

- ユーザからアミノ酸配列を受取り、相同タンパク質を検索後、ユーザがいくつかのタンパク質を選択し、マルチプルアラインメントを実行。その結果に基き、保存性が高い残基を表示。また、二次構造予測結果も同様にマルチプルアラインメントし、保存性が高い二次構造も表示。(1)、(2)同様、分散処理を行い、保存残基は立体構造が存在すればその上に表示。

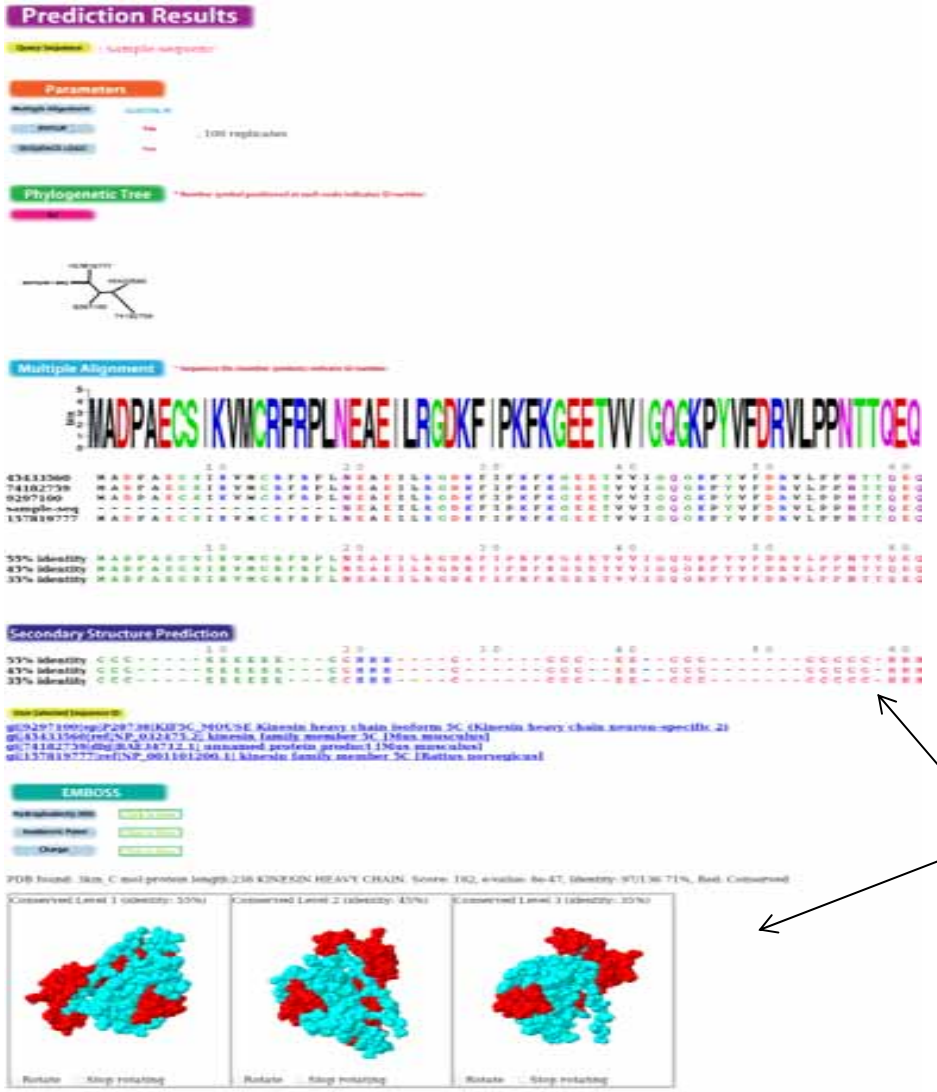
- 2008年3月31日より一般公開。

(3) タンパク質比較情報ワークフロー
(Protein Comparative Information Workflow)



(3) タンパク質比較情報ワークフロー
(Protein Comparative Information Workflow)

結果画面



保存度のレベルは3段階

2009年度計画

(1) タンパク質立体構造モデリングワークフロー

(Protein Structure Modelling Workflow)

2008**年度開発のワークフローと合わせ、タンパク質立体構造関連ワークフローファミリーの完成**

(2) アクティブ・ワークフローに向けた環境開発

(Active Workflow)

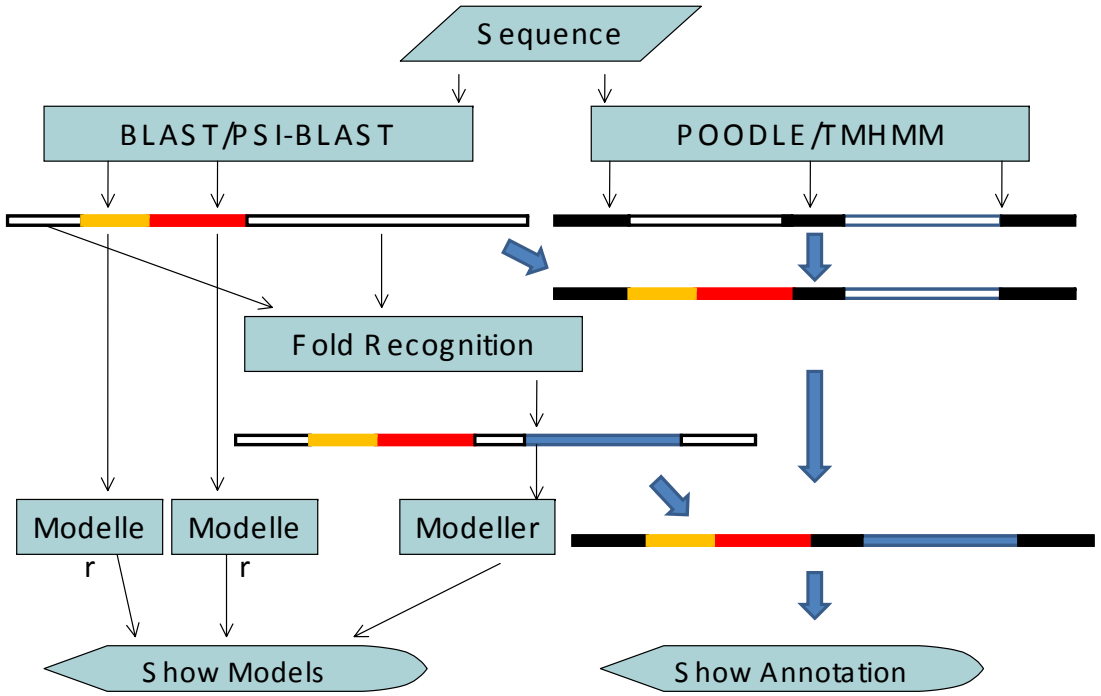
ユーザが定義可能または変更可能なプラットフォーム型ワークフローの構築

(1) タンパク質立体構造モデリングワークフロー (Protein Structure Modelling Workflow)

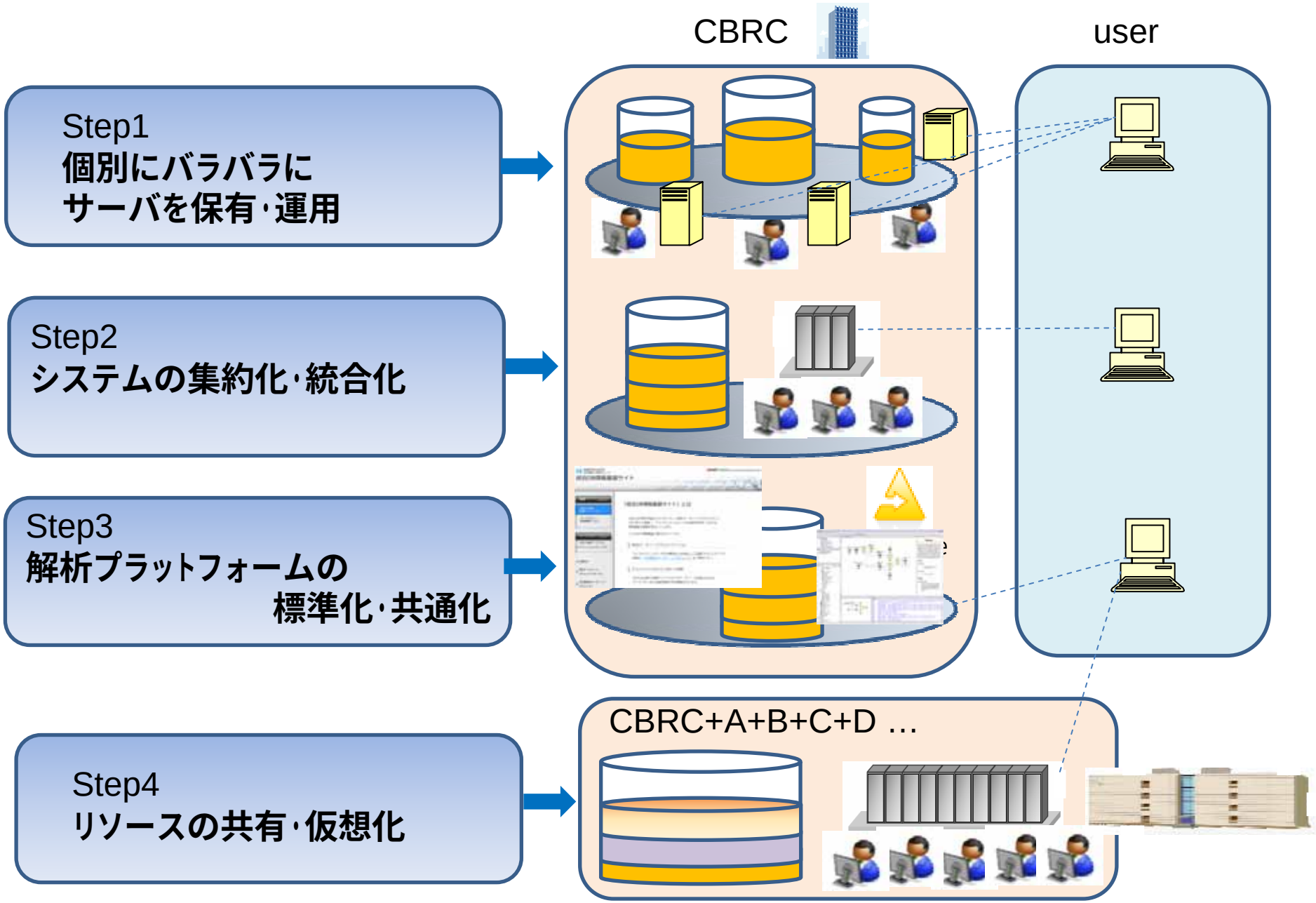
- 立体構造未知のタンパク質の構造モデルを実験研究者等に提供。

- ユーザからアミノ酸配列を受取り、BLAST/PSI-BLAST/構造認識法を基にドメイン分割し、ドメイン毎にモデル構築ツール(Modeller)を実行。
- さらに、POODLE/TMHMMなどタンパク質アノテーションワークフローから得られる情報を提供。

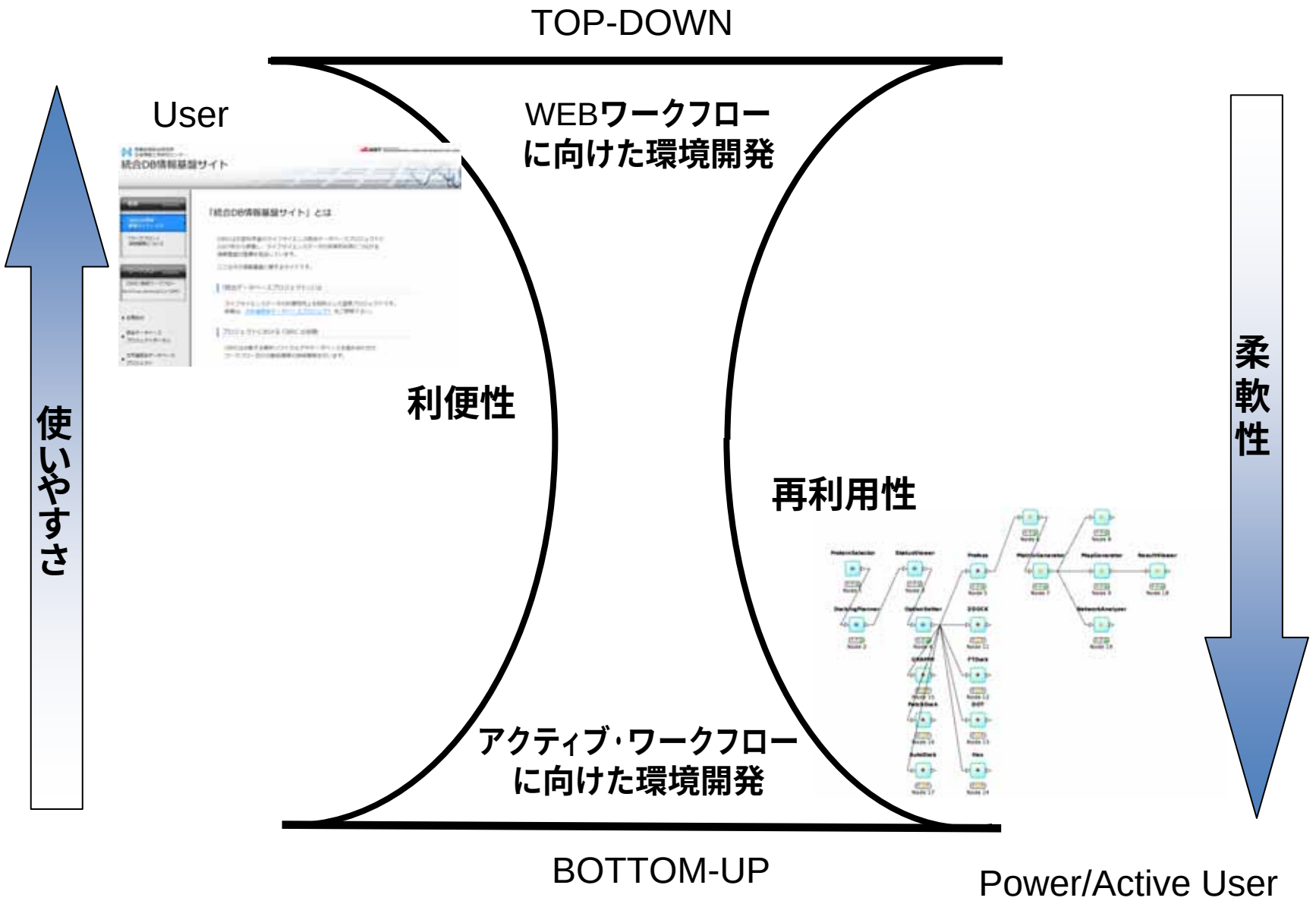
- 2009年12月末、一般公開予定。



CBRCワークフロー開発の流れ



CBRC:ワークフロー開発



ワークフロー・アプリケーション

主なワークフロー・アプリケーション

- Taverna (http://taverna.sourceforge.net)
- Kepler (https://kepler-project.org)
- KNIME (http://www.knime.org)

ワークフロー・アプリケーション比較

	軽快度	GUI	操作性	既存コンポーネント充実度	Webサービス対応	難易度
KNIME	◎	◎	◎	△	△	◎
Taverna	△	△	△	◎	◎	△
Kepler	△	○	△	△	○	○

操作性： KNIME ＞ Kepler ＞ Taverna

コンポーネント充実度： Taverna ＞ Kepler ≧ KNIME

ユーザに使っていただくためには？

敷居は低い方が良い

- ユーザフレンドリー
- アプリケーションの見た目

が重要ではないか？

エンドユーザを意識した場合、GUI、操作性の面でKNIMEが優位であると判断

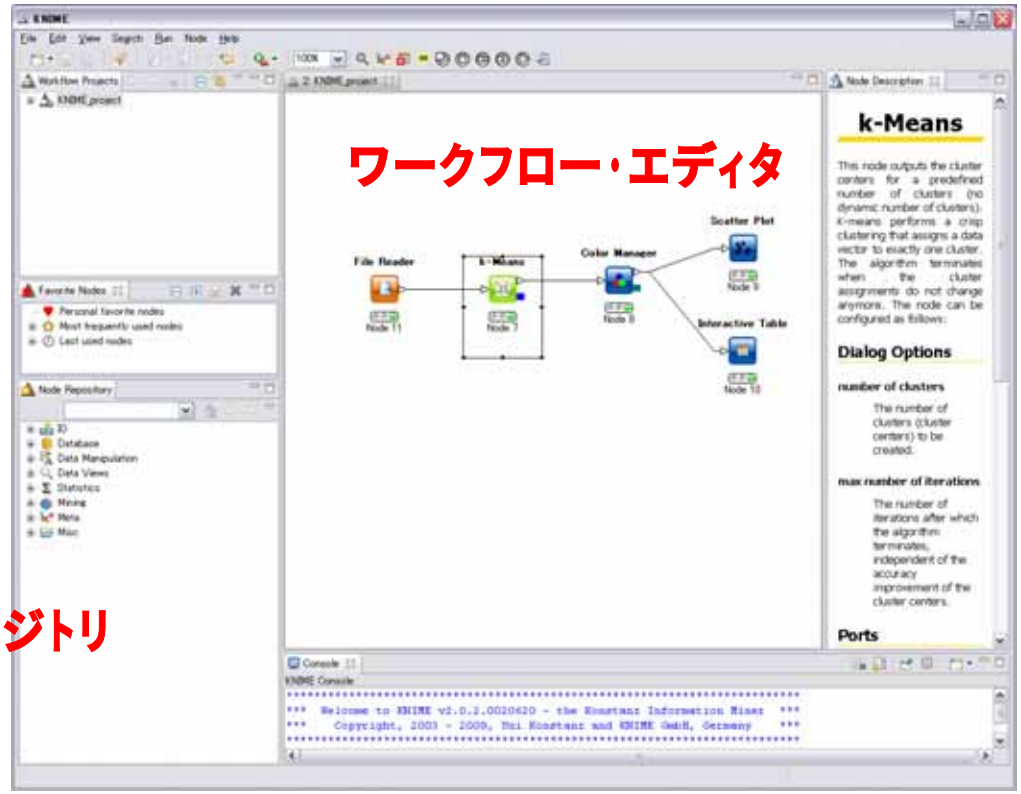
KNIME (Konstanz Information Miner)

メジャーなIDEであるEclipseをベースとしたプラットフォーム

お気に入り

ワークフロー・エディタ

ノード説明

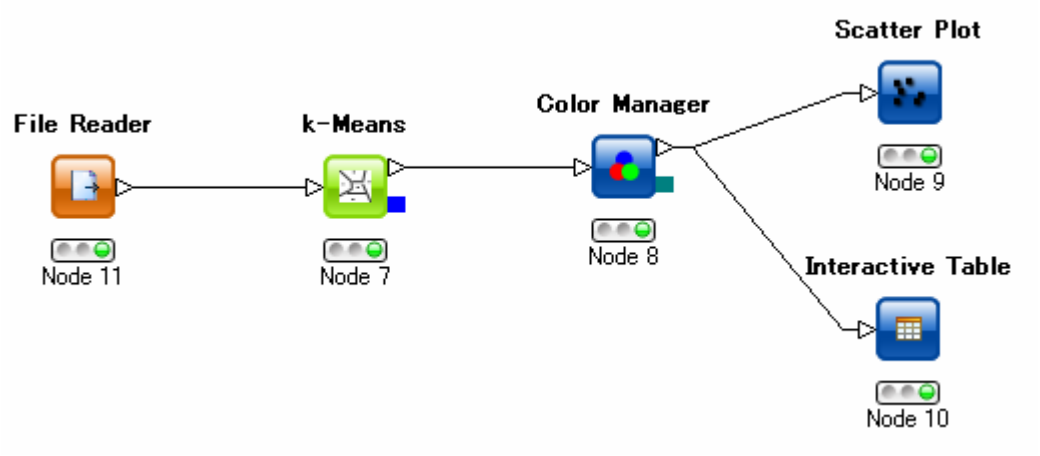


ノード・レポジトリ

- 見た目　→　直感的なGUI
- 操作性　→　ユーザフレンドリーな仕様
- 特筆すべき機能　→　対話的な実行、結果表示

ワークフロー・エディタ

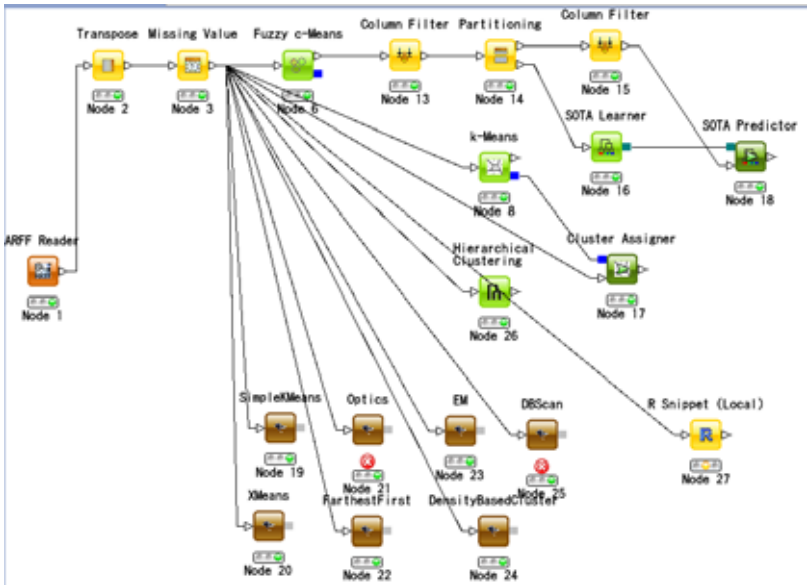
- ワークフローを構築するエディタ
- 操作性が容易
- アイコンレベルで実行状況を理解しやすい



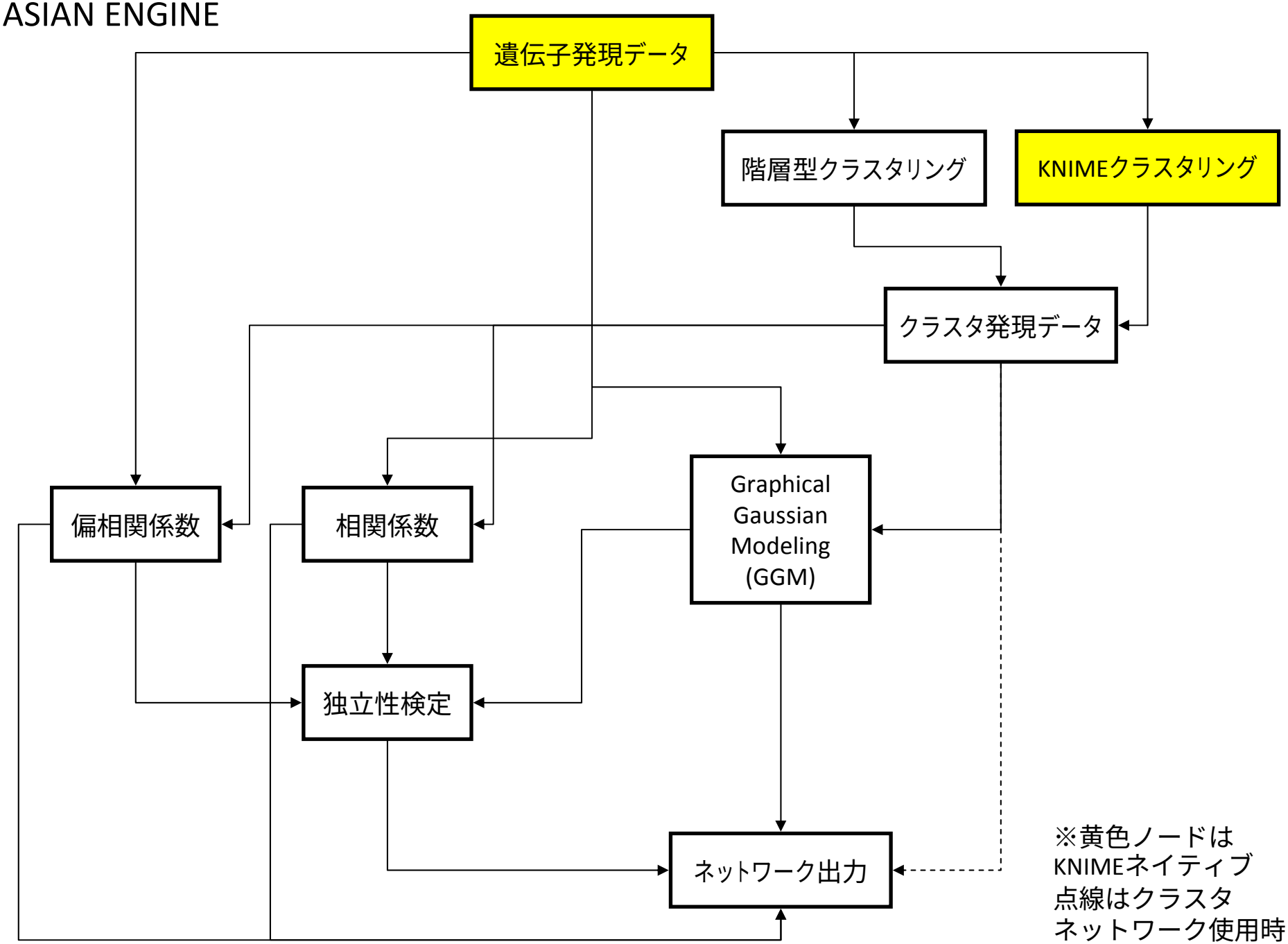
K-Means計算、描画ワークフロー

- 機能を視覚的に表現
- 実行状態をランプ表示
- 容易な各ノードの連結、移動
- 各ノードごとの実行も可能

ASIANワークフロー化
KNIMEによる基礎ノードの開発



ASIAN ENGINE



Knime.exe

WINDOWS版による プラットフォーム開発