D-033

メタデータを用いたバイオデータベース連携検索手法の提案

上田真由美* 高坂 貴弘† 細川 卓哉† 遠里由佳子† 伊達 進†
松田 秀雄† 下條 真司*

1 はじめに

今日，複数のインターネット上に分散した情報資源を共有することを可能にする，グリッドコンピューティングに関する研究が盛んになりつつある [1]．グリッドコンピューティングは，世界中に分散する異機種混合の計算環境を一つの巨大な仮想コンピュータとして扱うための技術である．したがって，巨大な計算機パワーやストレージを必要とする様々な分野での利用が求められている．

一方，バイオインフォマティックスの分野では，蛋白質の構造や機能解析，文献のテキストマイニングなど，膨大な量の情報と計算機パワーが必要とされている．さらに，塩基配列，アミノ酸配列，蛋白質立体構造など蓄積情報は飛躍的に増大し，様々なデータベースが構築されている．これらのデータベースは，それぞれ特徴があり，生物学の研究者は一つの情報を得るために，複数のデータベースに対して検索を行う必要がある．

本稿では，グリッド技術のデータベース検索への利用を目指し，メタデータを用いたバイオデータベースの連携検索手法を提案する．本手法の利用により，生物学の研究者は，インターネット上に分散するデータベースの所在・特徴を意識せずに，必要とする情報を容易に入手することが可能となる．

2 連携検索機構

本章では，バイオデータベース利用の現状と本システムに求められる要件について述べる．

2.1 バイオデータベース利用の現状

近年の分子生物学の発達により，塩基配列や蛋白質，文献データなど，大量の情報がデータベースに蓄積されている．これらのデータベースは，独自の目的に応じて構築されているため，類似した情報が格納されているのにも関わらず，異なる機能を持つ．したがって，生物学者や創薬関係者は，それぞれのデータベースの特徴と所在を的確に把握しておき，必要に応じて複数のデータベースを渡り歩いて検索する必要がある．さらに，検索結果の中から利用者の経験や知識，文献参照により絞込みを行い，他のデータベースの検索キーワードとして用いる．すなわち，現状では，利用者の経験と勘が非常に重要な位置を占めている．

2.2 検索機構

2.1 節のような状況を解決するため，様々な取り組みが行われている [2]．しかし，(1) 各データベースは独自に運営されており，(2) データベースとアプリケーションが 1 対 1 の対応ではない．(3) データベースは日々更新される．など理由から，データベースに変更を加え，1 つのデータベースに統合することは不可能である．

本研究では，各データベースとアプリケーションの間にメタデータの利用を検討する．

3 基本構成

本章では，提案手法を実現する基本構成について述べる．

3.1 構成

提案手法の構成を図 1 に示す．本システムは，既存のデータベースに対するアクセスに，OGSA-DAI[3] に含まれる GDSF(GridDataServiceFactory) と GDS(GridDataService) を用いる．OGSA-DAI は，OGSA に基づいて構築されたグリッド上でデータベースを利用するためのツール群である．GDSF が，ク

[1] 大阪大学サイバーメディアセンター
[2] 大阪大学大学院情報科学研究科

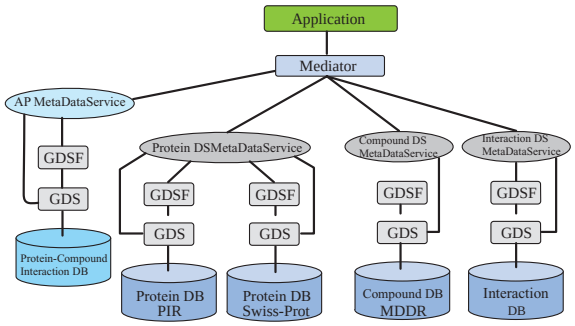

図 1: 連携検索機構の構成

表 1: 蛋白質 DS メタデータの例

蛋白質 DS メタデータ ID	SP ID	PIR ID
PDS00001	P12345	S98765
PDS00002	P23456	S87654

ライアントの要求に応じて GDS を生成することにより，様々なデータベースにアクセスすることが可能となる．

ここでは，蛋白質データベースとして Swiss-Prot[4] と PIR[5] を，化合物データベースとして MDL 社 [6] の MDDR を用いた．また，相互作用情報については，活性・抑制など相互作用の名称を持つ．

3.2 メタデータサービス群

本機構では，DS メタデータと AP メタデータを用いることにより，異種データベースの連携検索を実現する．さらに，本手法により，独自に様々な付加情報を追加することも可能となる．

- DS メタデータ
  DS メタデータは，対象毎にデータベースをグループ化し，グループ内の DB の異種性を解消し，統一的に扱うことを可能とする．DS メタデータは，各データベース内における ID の対応関係を用いて，DS メタデータ ID を与える（表 1）．

- AP メタデータ
  AP メタデータは，各グループ内の情報の関連性を保持する．すなわち，既知の情報から，ある蛋白質 DS メタデータ α とある化合物 DS メタデータ β の関連性 γ の組を持つ（表 2）．

表 2: AP メタデータの例

蛋白質 DS メタデータ ID	化合物 DS メタデータ ID	相互作用 DS メタデータ ID
PDS00001	CDS99999	ID00002
PDS00002	CDS77777	ID00005

4 まとめ

本稿では，インターネット上に多数存在するバイオインフォマティックス関連データベースの OGSA-DAI を用いた連携検索手法について提案した．本手法により，各データベースから独立したサービスの実現を可能とする．

今後の課題として，アプリケーションと各サービス間にメディエータを組み込むことにより，個人によるデータベースの利用傾向を反映したサービスや，実験より得た個人の持つデータベースの利用を試みる．さらにテキストマイニング技術の組み込みを考慮する．

謝辞

本研究は，文部科学省科学技術振興費主要 5 分野の研究開発委託事業の IT プログラム「スーパーコンピュータネットワークの構築」の助成を受けて行われた．バイオグリッドプロジェクトの諸氏に，大変貴重なご意見をいただいた．ここに記して感謝する．

参考文献

[1] グリッドコンピューティング特集, 情報処理学会誌, Vol.44, No.6, 2003.

[2] Integrating Biological Databases, Nature Reviews, Volume 4, pp.337–345, 2003.

[3] OGSA-DAI Project, http://www.ogsadai.org

[4] Swiss-Prot Protein knowledgebase, http://kr.expasy.org/sprot/

[5] Protein Information Resource, http://www-nbrf.georgetown.edu/pir/

[6] 日本 MDL, http://www.mdli.com/japan/