BioGrid：データグリッド

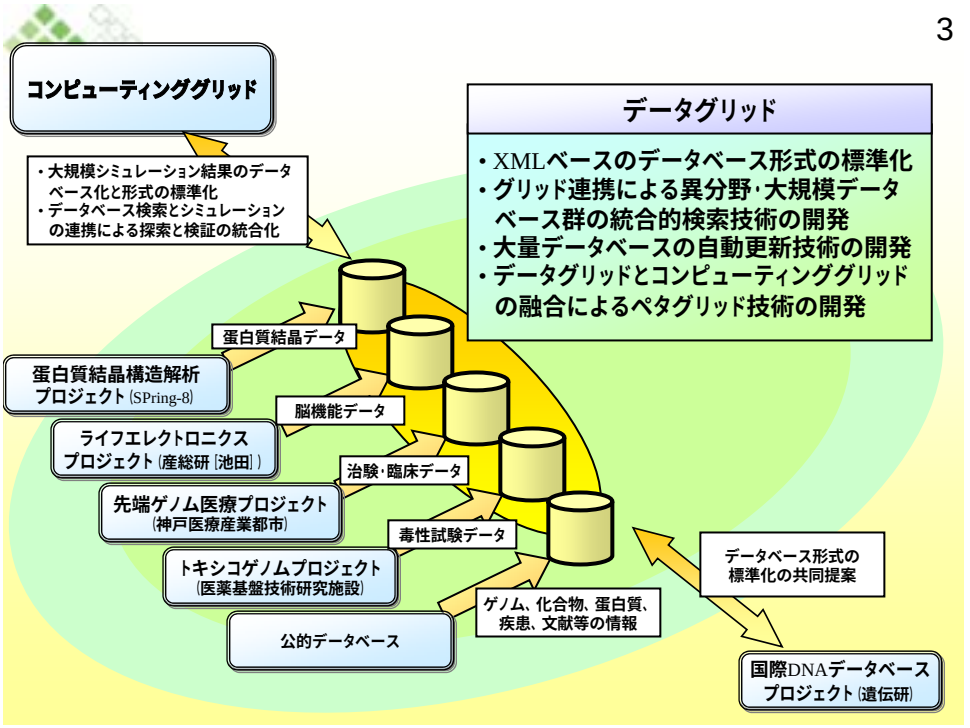
グリッド技術による異種データベース間の連携

松田秀雄
大阪大学大学院情報科学研究科
バイオ情報工学専攻
E-mail: matsuda@ist.osaka-u.ac.jp

背景

- インターネット上で多数のデータベースが散在
- (例) バイオ関連データベース：500個以上のデータベースが存在し、さらに数が増えつつある
- 検索のときの問題点
  - データベースのサービスとロケーション (どこにどのような種類のデータがあるか)
  - データベースごとのフォーマットの異種性 (同じ種類のデータでもデータベースごとに表記が異なる)
  - データベースの頻繁な更新と、量の急激な増大 (最新データの検索の必要性)
- 統合データベースの構築による解決には限界がある

⇨ グリッド技術によるデータベース連携

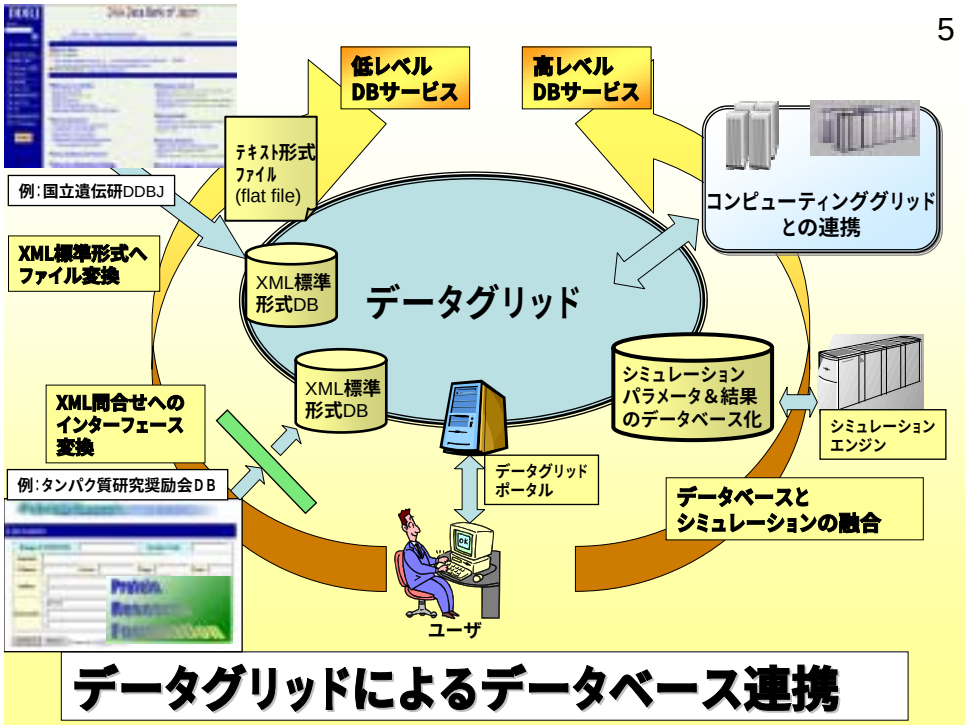


bioGrid

研究開発の目標

- データベースの異種性の解消（XML標準形式）
- グリッドデータベース基盤システムの開発
- ゲノム情報の検索：「全体」検索が必要（例 BLAST）

  →データベースのキャッシュ・自動更新
- 高次の情報検索：専門領域知識の利用／既存のウェブサービスとの連携



データベース書式の異種性

DNA 塩基配列 (DDBJ)

```
LOCUS       ECORBS 6197 bp DNA .
            30-OCT-1994
DEFINITION  E.coli ...rbsB ....
            transport system, ..
ACCESSION   M13169 M13517
NID         g147511
KEYWORDS    high affinity ribose
SOURCE      E.coli K12 DNA.
ORGANISM    Escherichia coli
REFERENCE   1  (sites)
  AUTHORS   ...
  TITLE     ...
  JOURNAL   ...
  MEDLINE   84032513FEATURES
gene            122..127
     /gene="rbsB"
     /translation="MKKGTVLNS...
                  ...
ORIGIN    94 bp upstream of BclI
          1 ctcaggttcg aaatctaac
```

タンパク質アミノ酸配列 (SWISS-PROT)

```
ID   RBSB_ECOLI ...
AC   P02925;
DT   21-JUL-1986 (...CREATED)
DT   21-JUL-1986 (... UPDATE)
DT   01-NOV-1995 (... UPDATE)
DE   D-RIBOSE-BINDING ...
GN   RBSB OR RBSP OR PRLB.
OS   ESCHERICHIA COLI.
OC   PROKARYOTA; GRACILICUTES;
OC   ENTEROBACTERIACEAE.
RN   [1]RP SEQUENCE FROM N.A.,...
RX   MEDLINE; 84032513.
RA   GROARKE J.N., MAHONEY W.C.,
RA   ZALKIN H., HERMODSON M.A.;
RL   J. BIOL. CHEM. 258:...
CC   -!- FUNCTION: INVOLVED IN
 ...
KW   TRANSPORT; SUGAR TRANSPORT;
KW   3D-STRUCTURE.
SQ   SEQUENCE   296 AA; ...
     MNMKKLATLV SAVALSATVS ANA...
     ...
```

タンパク質立体構造(PDB)

```
HEADER    SUGAR TRANSPORT 23-SEP-9
COMPND    D-RIBOSE-BINDING PROTEIN
COMPND    2 (G134R) COMPLEXED WITH
SOURCE    (ESCHERICHIA COLI)...
SOURCE    2 EXPRESSION PLASMID)
AUTHOR    S.L.MOWBRAY,A.J.BJORKMAN
REVDAT    1   26-JAN-95 1DRJ    0
JRNL        AUTH   A.J.BJORKMAN
JRNL        TITL   PROBING PR...
JRNL        TITL 2 RIBOSE-BINDING
JRNL        REF    TO BE PUBLISHED
JRNL        REFN   ASTM
SEQRES  1  271  LYS ASP THR ...
SEQRES  2  271  PRO PHE PHE ...
...
HELIX   1   A PRO     14  LEU ...
HELIX   2   B PRO     43  LEU ...
...
ATOM     1  N   LYS   1  x1 y1 z1
ATOM     2  CA  LYS   1  x2 y2 z2
...
```

XML標準形式作成のための変換テーブル

	Common Format	SWISS-PROT	PIR	PDB
エントリ	entry	entry	ProteinEntry	PDBj
エントリID	entry/id	entry/@name	ProteinEntry/@id	PDBj/@entry_ID
タンパク質名	entry/name	entry/proteinName	ProteinEntry/protein/name	PDBj/main/entity/entity_item/description
遺伝子名	entry/gene	entry/genes/gene/@name	ProteinEntry/genetics/gene/uid	PDBj/main/entity/entity_item/src_gen/gene
生物種（学名）	entry/organism/scientific	entry/organisms/organism/name	ProteinEntry/organism/formal	PDBj/main/entity/entity_item/src_gen/scientific_name
生物種（慣用名）	entry/organism/common	entry/organisms/organism/name	ProteinEntry/organism/common	PDBj/main/entity/entity_item/src_gen/common_name
文献	entry/reference	entry/references/reference	ProteinEntry/reference/refinfo	PDBj/main/citation/citation_item
機能	entry/function	entry/comments	---	---
EC#	entry/EC_num	entry/proteinName	---	PDBj/main/entity/entity_item/EC
キーワード	entry/keyword	entry/keywords/keyword	ProteinEntry/keywords/keyword	PDBj/main/struct/keywords
タンパク質配列	entry/sequence	entry/sequence	ProteinEntry/sequence	PDBj/main/entity/entity_item/seq_one_letter_code
タンパク質部分構造	entry/feature	entry/features/feature	ProteinEntry/feature	PDBj/struct/site
部分構造の型	entry/feature/type	entry/features/feature/@key	ProteinEntry/feature/feature-type	PDBj/struct/site/@id
部分構造の記述	entry/feature/description	entry/features/feature/@description	ProteinEntry/feature/description	PDBj/struct/site/site_gen/details

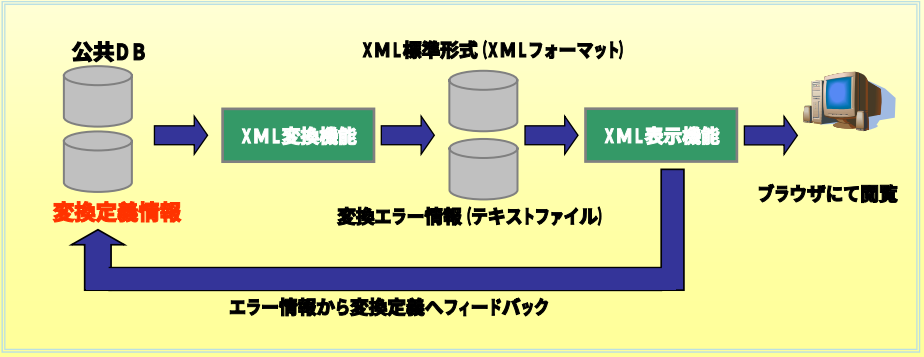
XML標準形式

```
<?xml version="1.0"?>
<entry>
  <id>データベースID</id>
  <date> 登録の日付 </date>
  <name db="pir" acc=".." >タンパク質名</name>
  <gene db="sp" acc="..">遺伝子名</gene>
  <organism db="pir" acc="..">
      <scientific db="pir" acc="..">生物種の学名</scientific>
      <common db="pir" acc="..">生物種の慣用名</common>
  </organism>
  <reference db="pir" acc="..">文献情報
     <author>著者</author> <citation>学術誌名</citation>  <volume>巻</volume>
     <year>発表年</year>      <title>文献題名</title>
     <first_page>開始ページ</first_page> <last_page>終了ページ</last_page>
  </reference>
  <function db="sp" acc=".."> <desc>タンパク質の機能記述</desc> </function>
  <keyword db="sp" acc="..">キーワードリスト</keyword>
  <feature db="pir" acc=".."> <type>部分構造のタイプ（活性部位など）</type> <desc>部分
       構造の説明</desc>
      <start>開始位置</start> <end>終了位置</end> </feature>
  <sequence db="sp" acc=".." length=".."> タンパク質の配列データ </sequence>
</entry>
```

阪大・情報科学　大川
アズテックシステム

公共データベースのXML変換システム

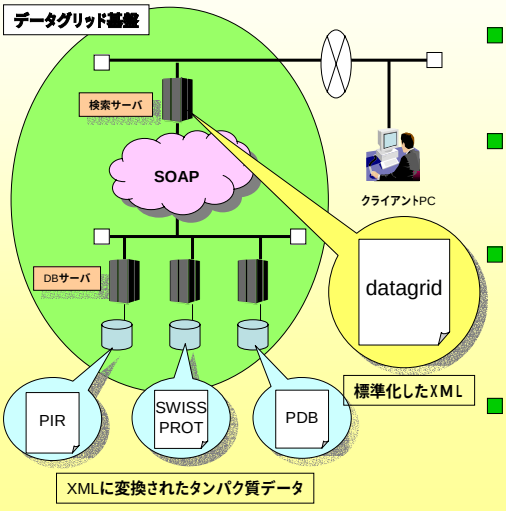
- XML標準形式変換機能
  ー公共データベースを変換定義規則に基きXML標準形式ファイルへ変換する。
- XML標準形式表示機能
  ー変換されたXML標準形式ファイルと対応する、公共データベース変換処理の際に発生したエラー
  内容を対応付けして表示する。



日立ソフト

グリッドデータベース基盤システムのプロトタイプ開発

ネットワーク上で広域分散したデータベース対してシームレスな検索が可能



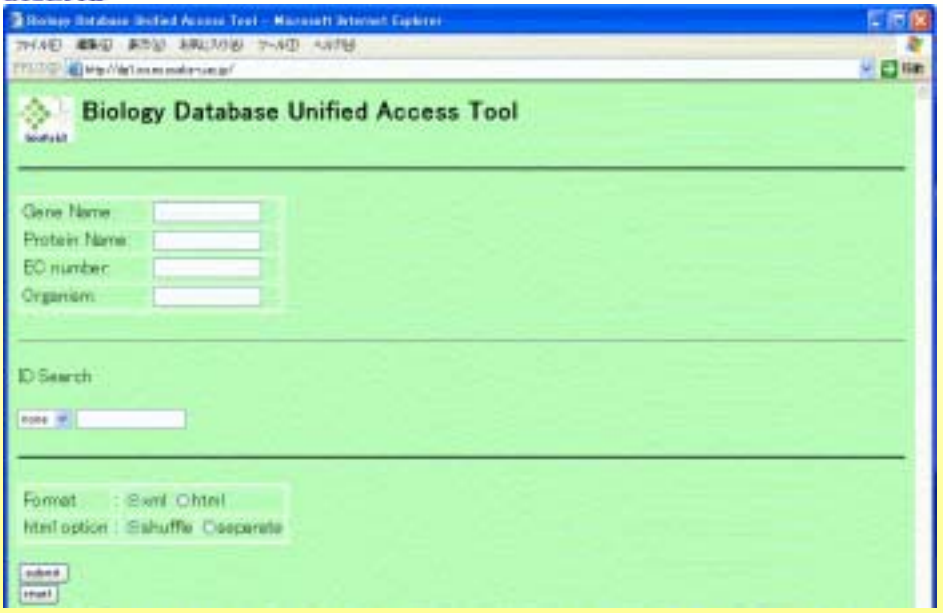
- スキーマの異なる複数のXMLデータを、標準化したXML形式のデータとして仮想的に統合
- ユーザに対してスキーマの違いを意識させない検索機能を提供
- 検索結果はXMLデータにあわせた表示形式に変換され、Webブラウザ上で閲覧可能
- データベースをSOAPを用いて接続することによって、広域分散したデータベースを利用可能

グリッドデータベース基盤システムのプロトタイプ開発（つづき）

トップ画面　　データ内容表示画面

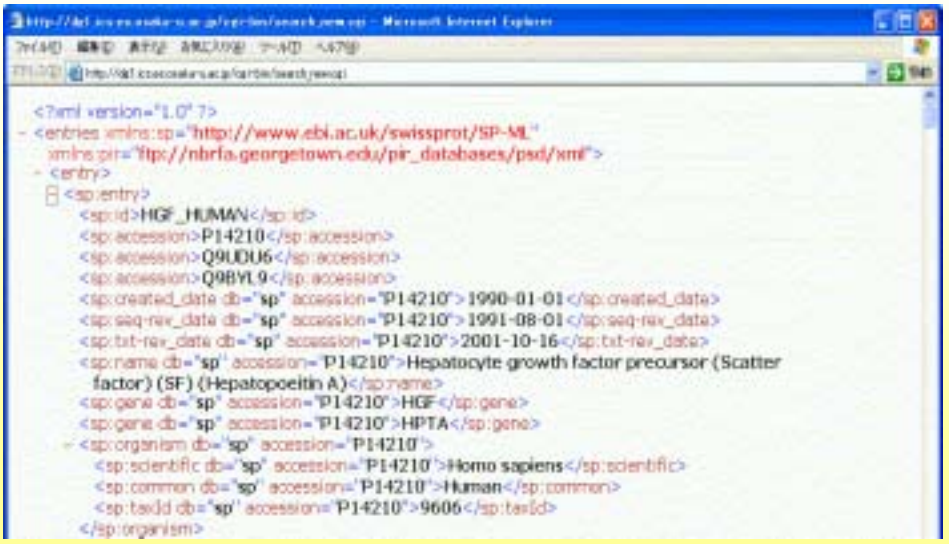
- 標準化XML形式の要素に対してキーワードサーチで検索
- グリッド上のSWISS PROT、PIR、PDBの各データベースを検索
- 問合せ結果のＸＭＬデータをＨＴＭＬに変換してブラウザに表示

データベース統合検索ツール



bioGrid

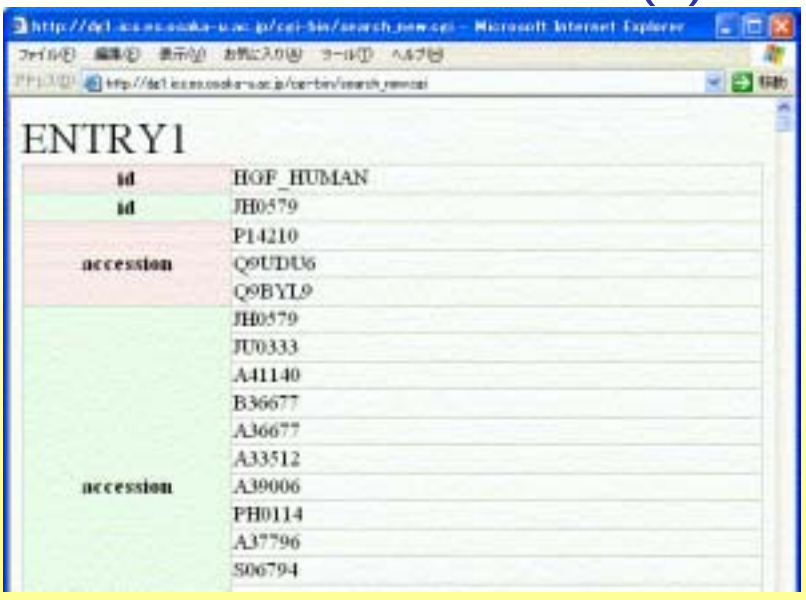
XML表示



```
<?xml version="1.0" ?>
- <entries xmlns:sp="http://www.ebi.ac.uk/swissprot/SP-ML"
  xmlns:pir="ftp://nbrfa.georgetown.edu/pir_databases/psd/xml">
  - <entry>
    - <sp:entry>
        <sp:id>HGF_HUMAN</sp:id>
        <sp:accession>P14210</sp:accession>
        <sp:accession>Q9UDU6</sp:accession>
        <sp:accession>Q9BYL9</sp:accession>
        <sp:created_date db="sp" accession="P14210">1990-01-01</sp:created_date>
        <sp:seq-rev_date db="sp" accession="P14210">1991-08-01</sp:seq-rev_date>
        <sp:txt-rev_date db="sp" accession="P14210">2001-10-16</sp:txt-rev_date>
        <sp:name db="sp" accession="P14210">Hepatocyte growth factor precursor (Scatter
          factor) (SF) (Hepatopoeitin A)</sp:name>
        <sp:gene db="sp" accession="P14210">HGF</sp:gene>
        <sp:gene db="sp" accession="P14210">HPTA</sp:gene>
      - <sp:organism db="sp" accession="P14210">
          <sp:scientific db="sp" accession="P14210">Homo sapiens</sp:scientific>
          <sp:common db="sp" accession="P14210">Human</sp:common>
          <sp:taxId db="sp" accession="P14210">9606</sp:taxId>
        </sp:organism>
```


bioGrid

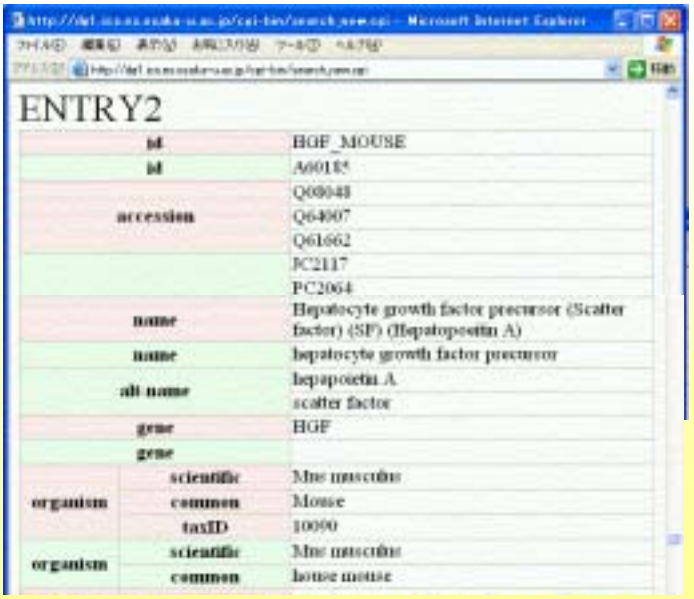
XSLTによるHTML表示(1)



ENTRY1

id	HGF_HUMAN
id	JH0579
accession	P14210
	Q9UDU6
	Q9BYL9
accession	JH0579
	JU0333
	A41140
	B36677
	A36677
	A33512
	A39006
	PH0114
	A37796
	S06794

XSLTによるHTML表示(2)



ENTRY2

id		HGF_MOUSE
id		A60185
accession		Q08048
		Q64007
		Q61662
		PC2117
		PC2064
name		Hepatocyte growth factor precursor (Scatter factor) (SF) (Hepatopoietin A)
name		hepatocyte growth factor precursor
alt name		hepapoietin A
		scatter factor
gene		HGF
gene		
organism	scientific	Mus musculus
	common	Mouse
	taxID	10090
organism	scientific	Mus musculus
	common	house mouse

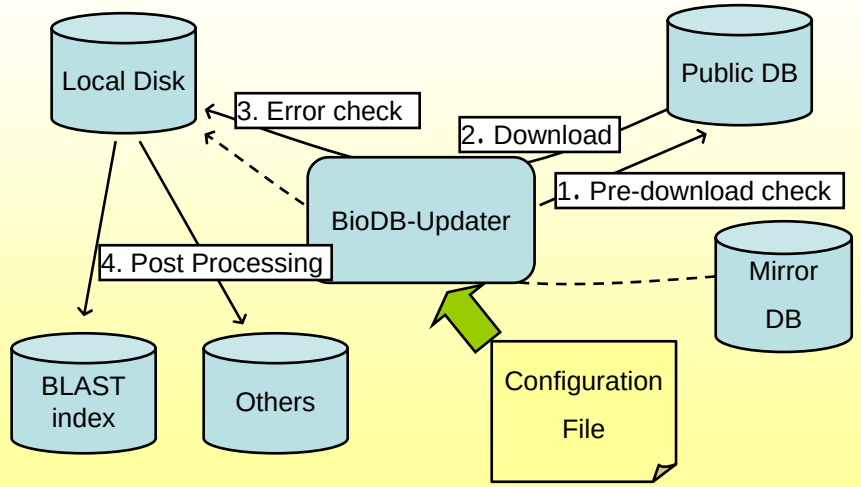
XSLTによるHTML表示(3)

function	description	SPECTRUM OF TISSUES AND CELL TYPES. IT HAS NO DETECTABLE PROTEASE ACTIVITY DIMER OF AN ALPHA CHAIN AND A BETA CHAIN LINKED BY A DISULFIDE BOND A SHORT FORM OF HGF IS PRODUCED BY ALTERNATIVE RNA SPLICING. THE SEQUENCE SHOWN HERE IS THAT OF THE LONG FORM CONTAINS 4 KRINGLE DOMAINS BELONGS TO PEPTIDASE FAMILY S1; ALSO KNOWN AS THE TRYPSIN FAMILY. PLASMINOGEN SUBFAMILY
	submit	DIMER OF AN ALPHA CHAIN AND A BETA CHAIN LINKED BY A DISULFIDE BOND
	alternative_products	A SHORT FORM OF HGF IS PRODUCED BY ALTERNATIVE RNA SPLICING. THE SEQUENCE SHOWN HERE IS THAT OF THE LONG FORM
	similarity	CONTAINS 4 KRINGLE DOMAINS
	similarity	BELONGS TO PEPTIDASE FAMILY S1; ALSO KNOWN AS THE TRYPSIN FAMILY. PLASMINOGEN SUBFAMILY
function	superfamily	SF001152
		hepatocyte growth factor
		kringle homology
		trypsin homology

公共データベースの自動更新システムの開発

- 更新の重要性
  - 遺伝子データベースは毎日平均1500万塩基の割合で増加
  - 最新データへのアクセスが研究の成否に直結
- BioGridプロジェクトにおける意義
  - プロジェクト遂行に必要な公共データベースの提供
  - 公共データベースを基にプロジェクトで開発される新規統合データベースの自動更新

データベース自動更新システムの動作



検索用配列分類データベースHoCDB

（Homology-based Clustering DataBase）

現在のバイオデータベースにおける問題点

バイオデータベースのサイズは指数関数的なペースで増大！
代表的な蛋白質データベース　SwissProt(+TrEMBL)のエントリ数は約85万

そのため、現状では検索効率が非常に悪い・・・

例えばBLASTによりデータベースを検索すると、その遺伝子のホモログが検索の上位を占めてしまって、本当に欲しい配列を見つけるのに非常に手間がかかる

解決策

1．冗長性の削減

データベースの検索結果を動的に階層化することにより、検索結果の冗長性を削減する

HoCDBでは、配列の相同性に基づいて、冗長な配列同士を1つにまとめて表示します

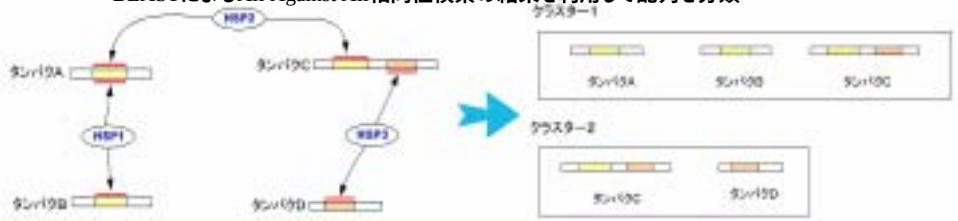
2．データの味付け

他のバイオデータベースの情報（2次情報）を出力結果に添える（データの味付け）ことによって、より精度の高い、情報のブラウジングが可能になります

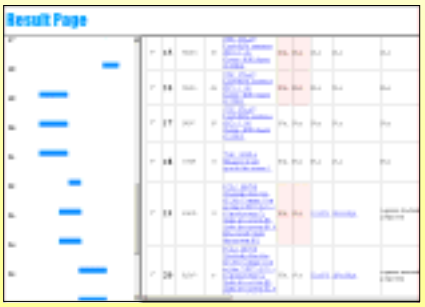
HoCDBでは、Swiss-Protのコメント、PDBの立体構造情報、SCOPの立体構造分類情報など、蛋白質に関する様々な2次情報を表示させることができます

配列分類アルゴリズム

BLASTによるAll Against All相同性検索の結果を利用して配列を分類



検索結果表示画面



クラスターメンバー表示画面

高次の情報検索システム：ＡＤＭＥ（薬物の吸収・分布・代謝・排出）情報XMLデータベースシステムの概要

(1) ＡＤＭＥ情報XML形式の設計

・ＡＤＭＥ情報標準化XMLスキーマ設計

(2) ＡＤＭＥ情報XML変換プログラムの開発

・XMLデータベースへの一括登録機能の開発

(3) ＡＤＭＥ情報XML検索システムの開発

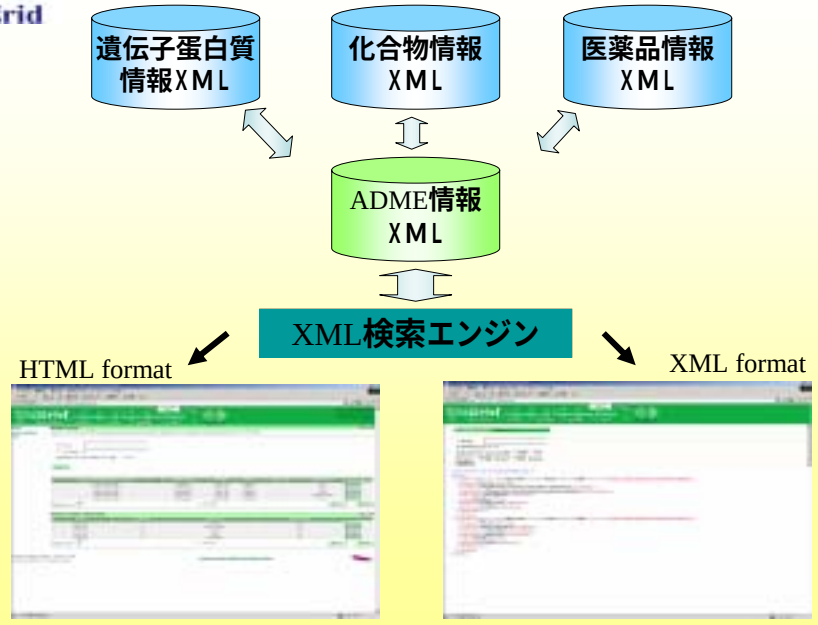
・インプット/アウトプット形式の検討・実装

・ＡＤＭＥ情報関連XML(遺伝子蛋白質情報XML、化合物情報XML 医薬品情報XML)とのリンク機能の検討

(4)表示インターフェイスの開発

・ADME情報及びADME関連情報(遺伝子蛋白質情報、化合物情報、医薬品情報)の表示方式の検討/開発

ADME情報XMLデータベースの実行イメージ



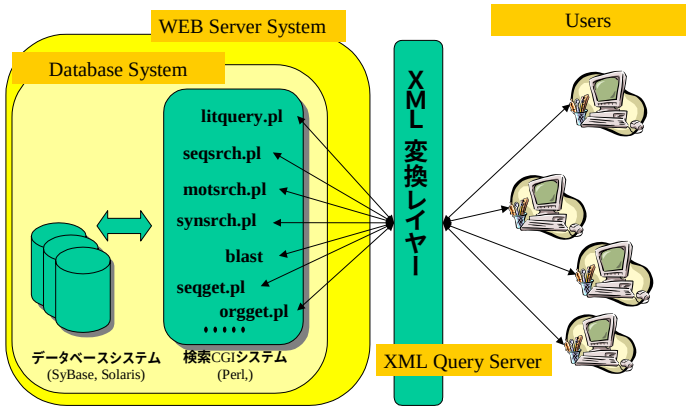
既存のウェブサービスとの連携：XML変換レイヤーの設計

1. 実際に運用されているデータベースシステムに対して、最小限のカスタマイズで仮想的統合を実現することができる。
2. データベースのサービス提供側のメリット：
   - データベースの全データのコピーを提供しなくてすむ→コピーの更新を考える必要がない。
   - 部分的なアクセス制限をかけてサービスのレベル分けができる。

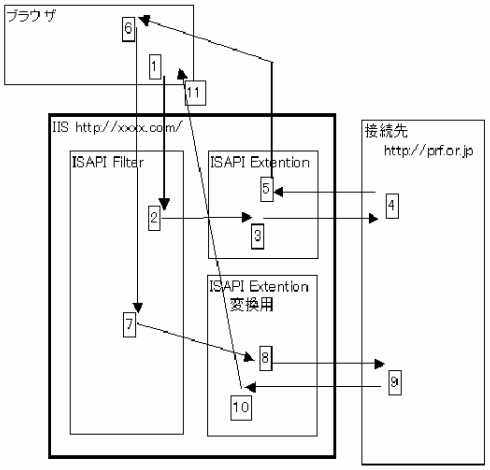
→ビジネス化にも向く

蛋白質研究奨励会 (PRF) データベースのサービス

PRFデータベースのXML問合せシステム



XML変換レイヤーの動作構成



1 ブラウザから、HTMLファイルをリクエスト
2 登録してあるサーバーのURLに変換
3 Extention から外部のサーバーにアクセス
4 外部のサーバーがHTMLを返す
5 そのまま、ブラウザに返信する
6 ブラウザから、CGIをリクエスト
7 変換用のExtentionにリクエストを送る
8 外部のサーバーへCGIリクエストを出す
9 HTMLを返す
10 HTMLをXMLへ変換する
11 ブラウザへ返す ここで、XSLT 変換

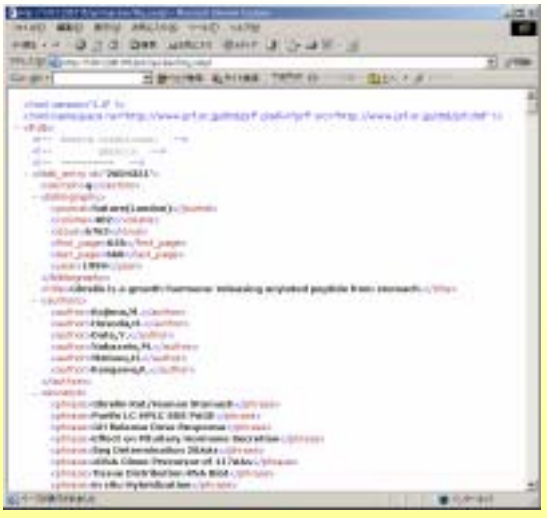
動作の実際 (データベースサービス)

XMLレイヤーを介したHTTPリクエスト

動作の実際（データベースサービス）

検索CGIシステムのリクエストと検索結果の表示（HTML版）

動作の実際（データベースサービス）



検索CGIシステムのリクエストと検索結果の表示（XML版）

まとめ

- 今年度の成果
  - データベースの異種性の解消
  - データグリッドの要素システム開発（DBの自動更新、配列分類、XMLベース検索、専門知識の体系化）
- 今後の研究開発
  - データグリッド基盤システムの完成
  - さらに多様なデータベースの結合（ゲノム、生体ネットワークなど）
  - コンピューティンググリッドとの連携（コンピューティンググリッドの結果をデータベース化）
- 成果発表
  - 日本分子生物学会（横浜） 2002.12.11-14
  - GlobusWorld(SanDiego) 2003.1.13-17 Life Science Workshop
  - PRAGMA(福岡） 2003.1.23-24 Life Science Workshop