「次世代人工知能・ロボット中核技術開発」

（人工知能分野）中間成果発表会

－人間と相互理解できる人工知能に向けて－

酵素反応データベースに向けた文献キュレーション支援技術の研究開発

平成29年３月29日

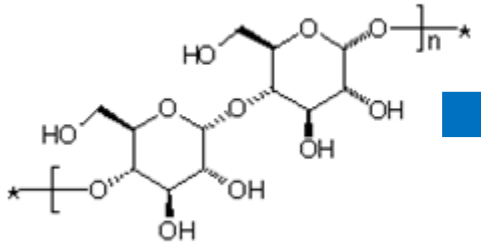
国立研究開発法人 産業技術総合研究所

長野 希美

国立研究開発法人 産業技術総合研究所

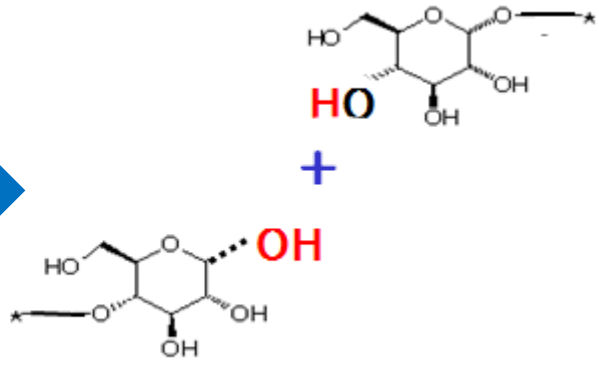
国立研究開発法人 新エネルギー・産業技術総合開発機構

自然界における多様な酵素



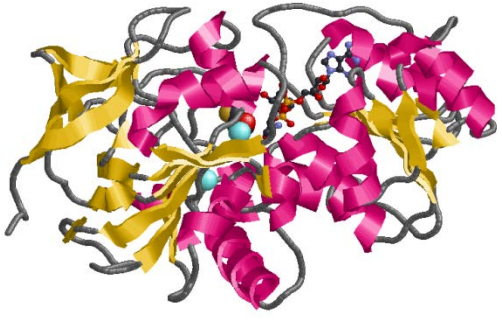
デンプン（スターチ）

唾液等に含まれるαアミラーゼ



小分子へと分解されたデンプン

アルコール
（エタノール）



アルコールデヒドロゲナーゼ

アセト
アルデヒド

アルデヒドデヒドロゲナーゼ

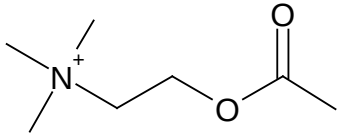
酢酸

肝機能指標；
γ-GTP

γグルタミルトランスペプチダーゼ
の活性が指標

自然界には、5千種類以上の酵素があるとされ、あらゆる反応を担っている。

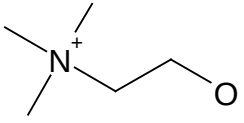
薬物の作用機序と酵素反応機構



神経伝達物質；
アセチルコリン

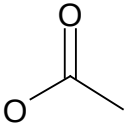
アセチルコリンエステラーゼ
（AChE）

アセチルコリンを分解



コリン

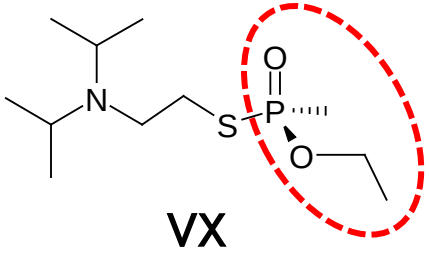
+



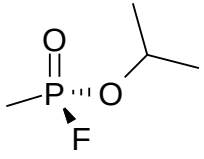
酢酸

薬物の作用機序と酵素反応機構

神経ガス

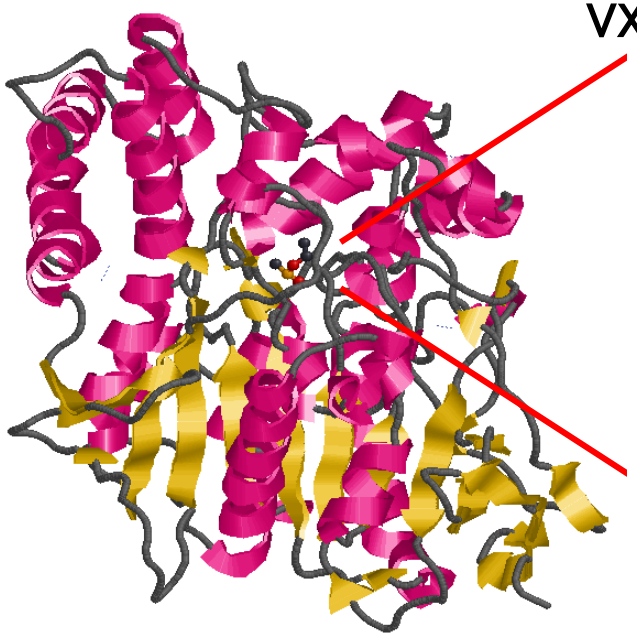


VX



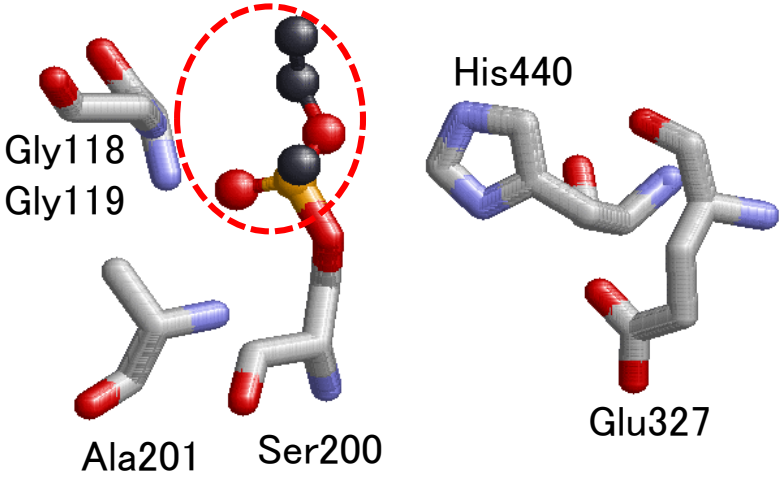
サリン

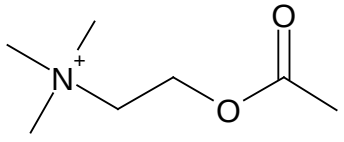
アセチルコリンエステラーゼ
（AChE）



AChEの活性部位

VXの一部が活性部位に結合して外れない





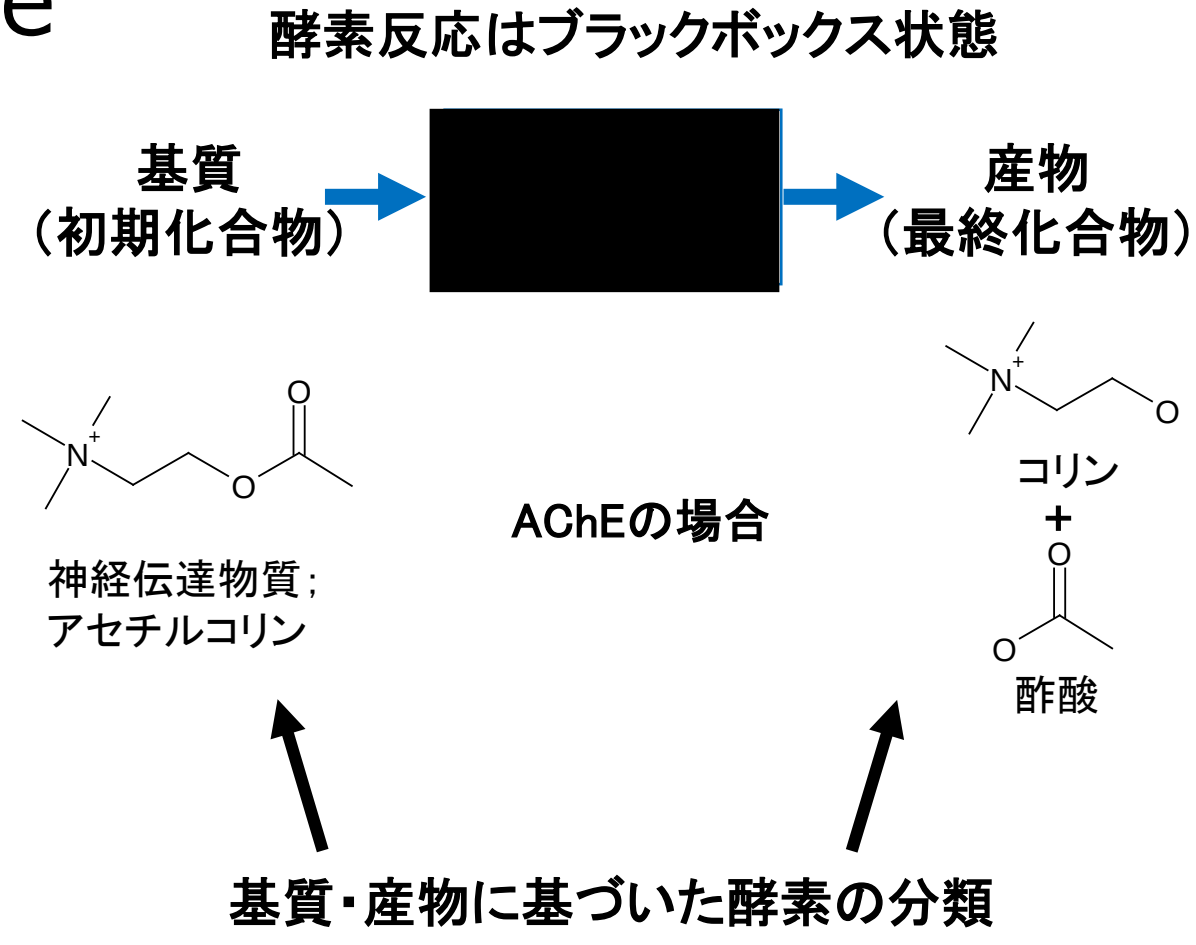
本来、分解すべき、アセチルコリンが結合できず、アセチルコリンを分解できなくなる

毒物、薬物の作用機序を解明する際には、酵素の反応機構が重要

酵素データベースの問題点

- 数々の酵素DB
  - IUBMB-Enzyme nomenclature
  - IntEnz
  - BRENDA
  - KEGG
  - ExPASy
  - …

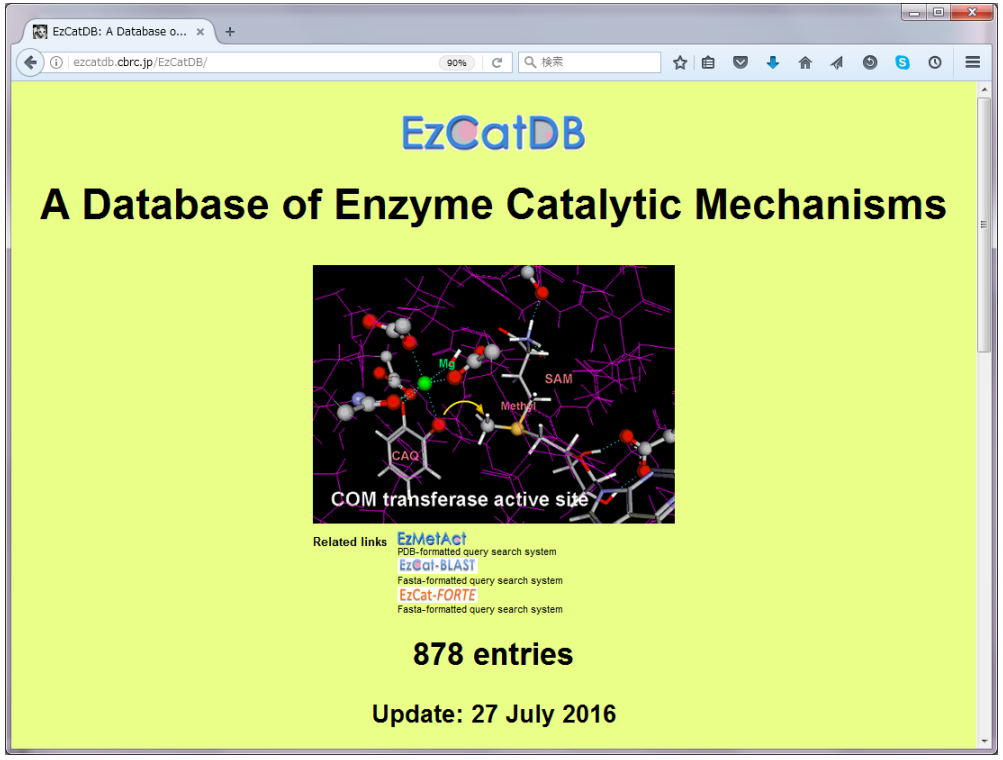
従来の酵素分類の問題点



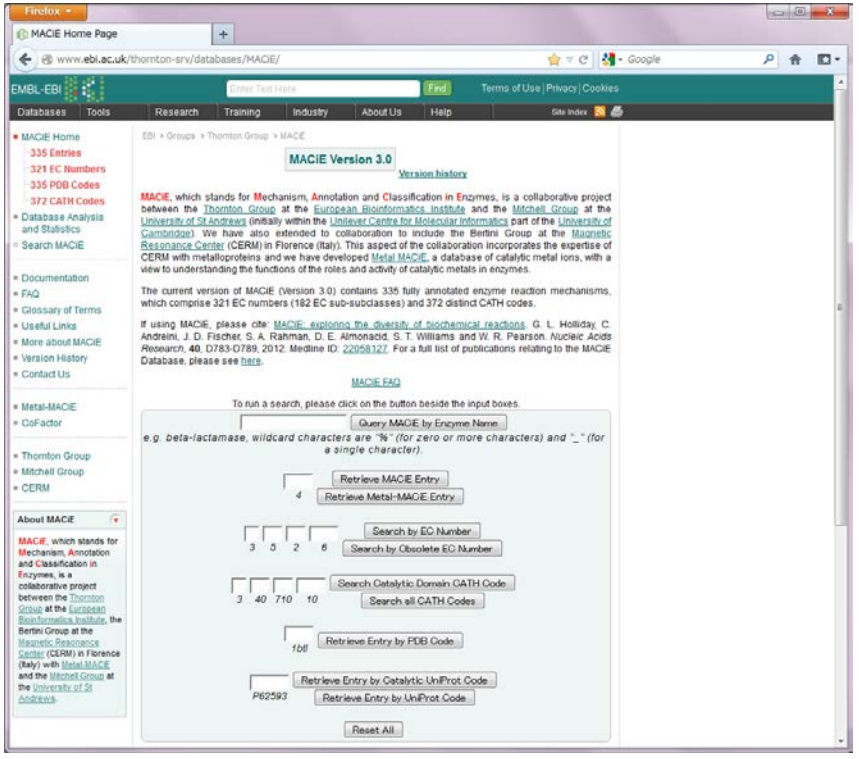
酵素反応データベースの問題点

- 酵素反応データベース

産総研；EzCatDB



欧州バイオインフォマティクス研究所；MACiE



酵素エントリ数；878

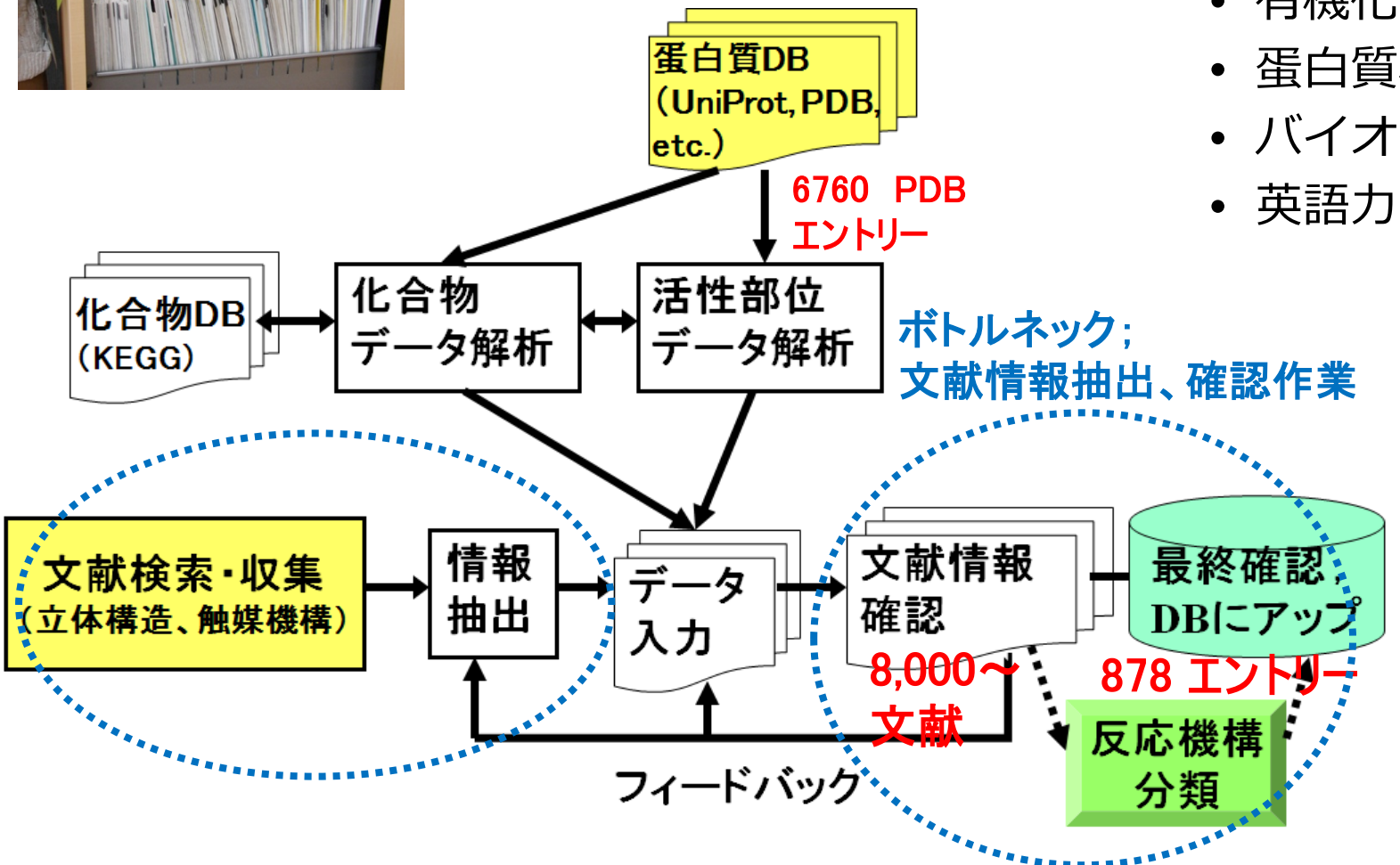
酵素エントリ数；335

- **酵素反応機構は重要にも関わらず、酵素反応データベースは数少ない。**
- **数千の酵素に対して、エントリ数も数百程度と少ない。**

酵素反応データベースの問題点

- EzCatDBのデータ作成スキームと必要なスキル

酵素の文献全体；300万件
収集文献；20,000～



- 必要なスキル
  - 生化学
  - 有機化学
  - 蛋白質科学
  - バイオインフォマティクス
  - 英語力

キュレータの
人材不足

約15年で、
約8000文献の
情報解析

文献解析の
効率化が必要

酵素反応データベースと自然言語処理とのコラボ

英国マンチェスター大学
テキストマイニング・システム

膨大な酵素関連
英語文献情報

情報抽出

**新たに
酵素反応分野に
適用**

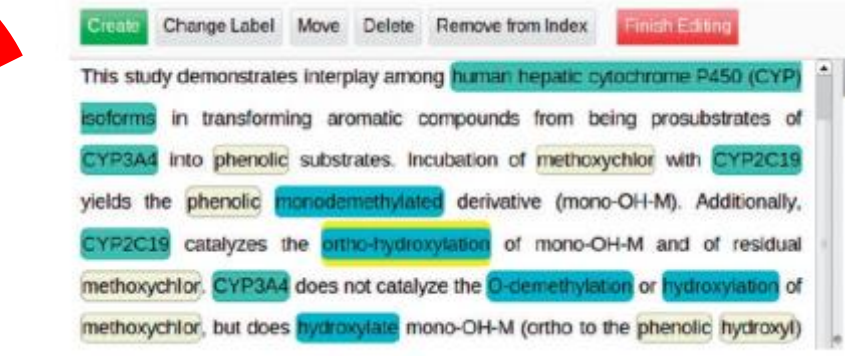
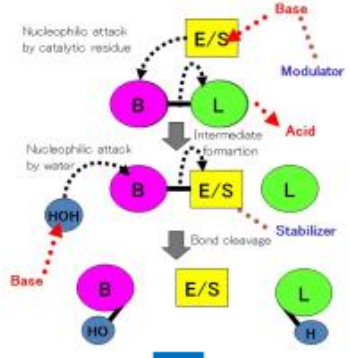


Fig.3 of Rak, *et al.*, (2014) *Database (Oxford)* **2014**, pii:bau070.

複雑な酵素反応機構

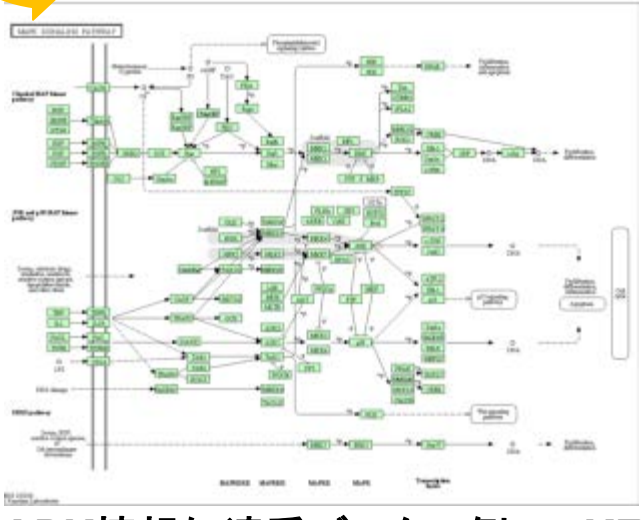


分類・登録

酵素反応データベース・EzCatDB

酵素反応に関する
専門知識をAIに移す
**“AI for 科学技術研究”
の実例**

**情報伝達系、代謝経路のキュレーション
に適用実施**

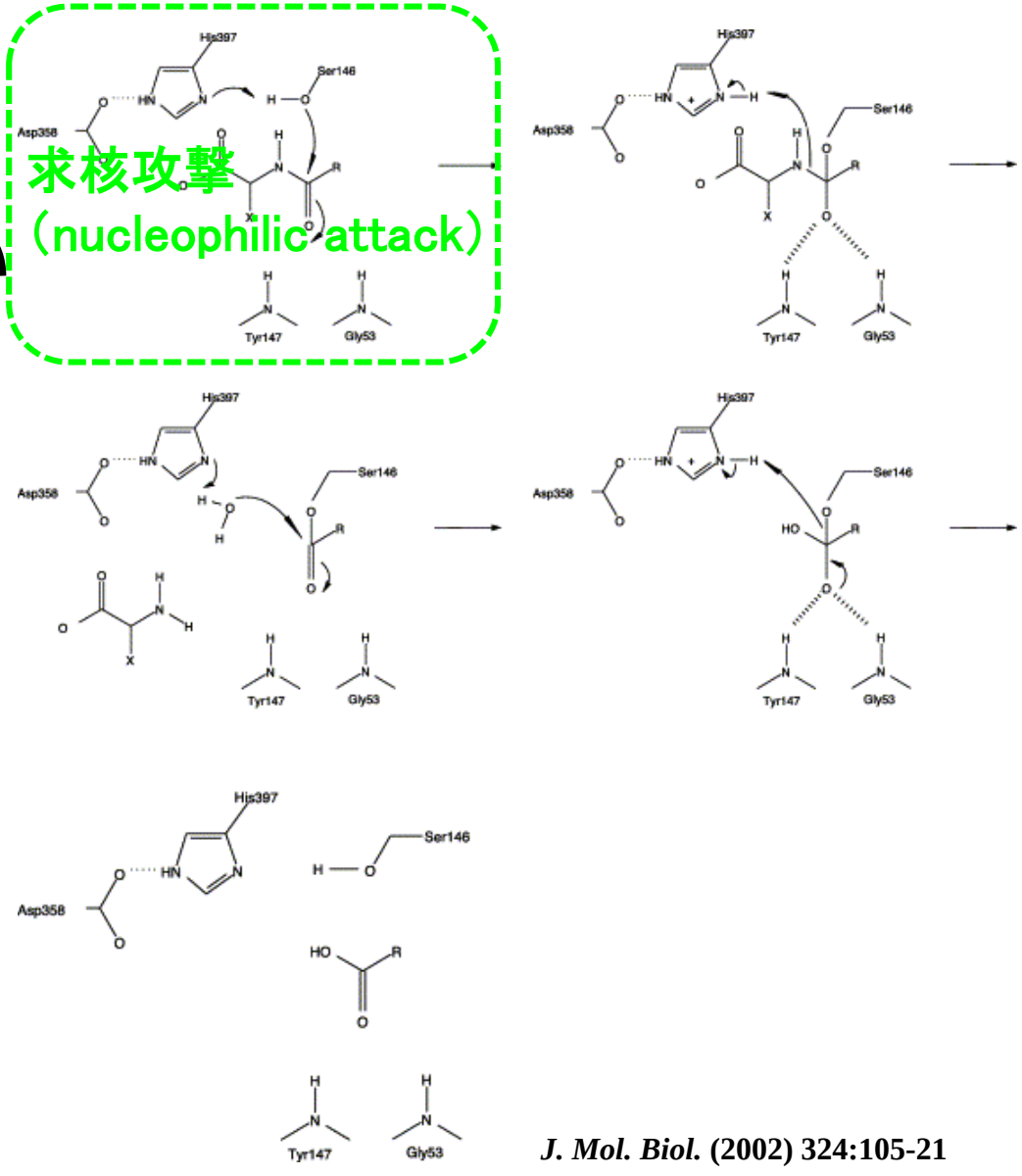


MAPK情報伝達系データの例　@KEGG

複雑な酵素反応機構の文章表現

- 文献における反応機構の表現
  - 図で表現→理解しやすい
  - 多くは、文章で表現
- 文章表現の問題
  - 同一反応；複数の表現

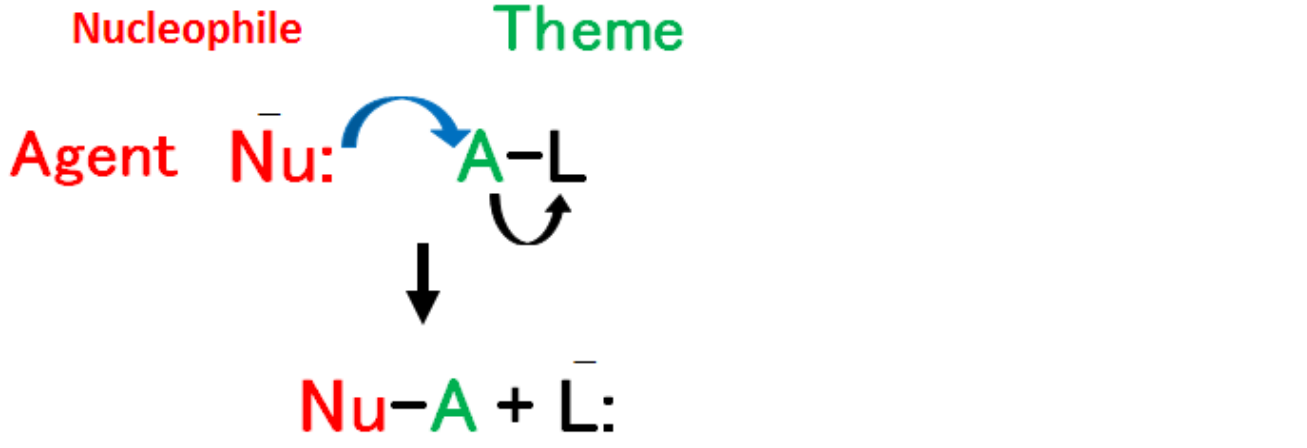
セリンプロテアーゼの反応機構図



J. Mol. Biol. (2002) 324:105-21

反応事象の文章表現

反応事象(Event type)；求核攻撃(nucleophilic attack)



Agent（主語） Event type（述語） Theme（目的語）

NuがAを求核攻撃する

Active-site-Nu acts as a nucleophile to attack on A
= Active-site-Nu makes a nucleophilic attack on A
= (Nucleophilic) active-site-Nu attacks A
=A is attacked by active-site Nu

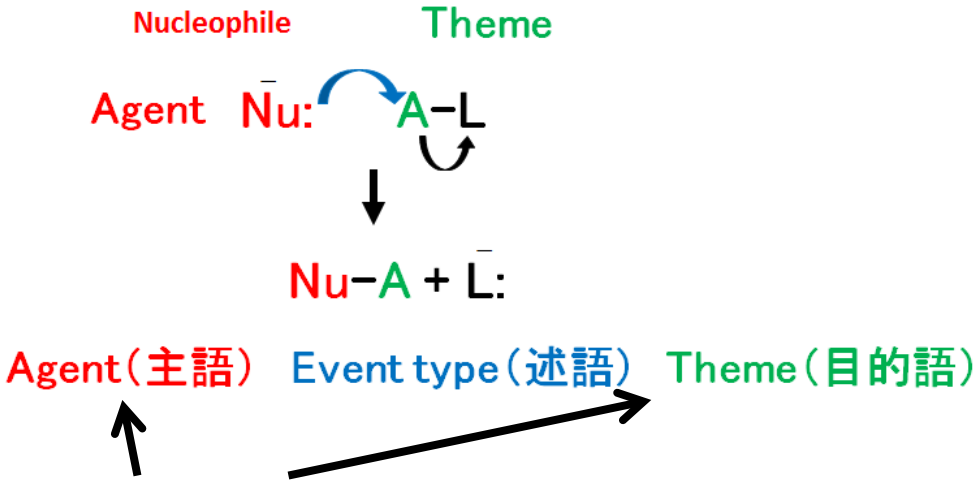
著者、論文によって異なる表現；
同一事象（イベント）に複数の表現

複数の表現を統一して標準化して欲しい ⟷ **同一表現だと剽窃を疑われる（？）**

事象表現の標準化は難しい！！
古い文献には対応できない！！

➡ **AI(テキストマイニング)による学習**

反応を担う実体（エンティティ）の複数表現

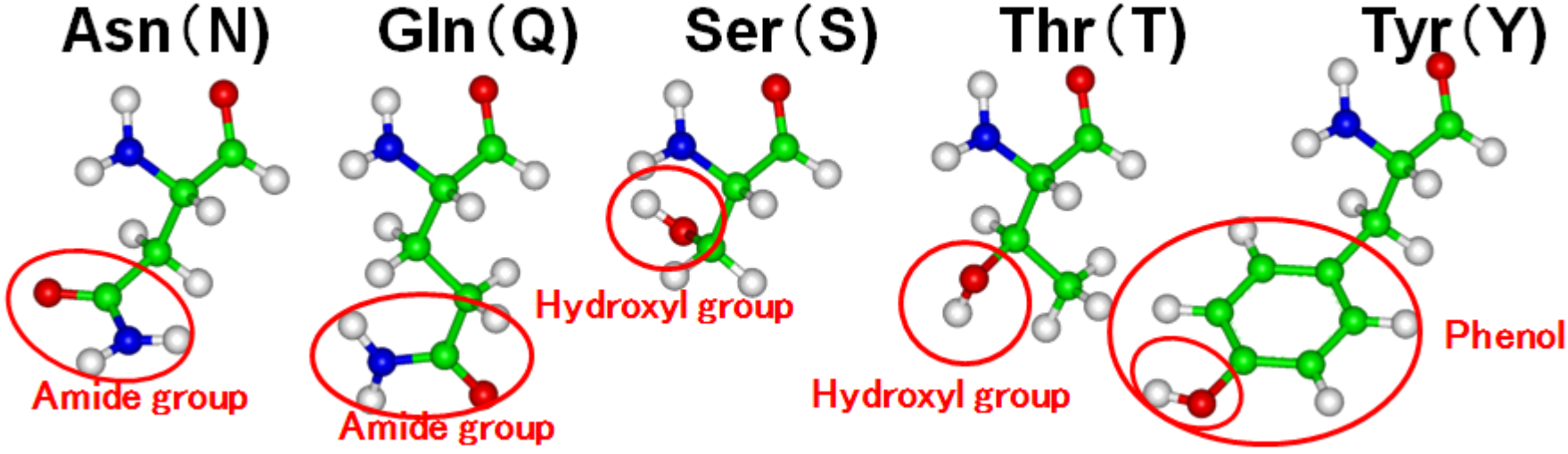


主語、目的語となる実体（エンティティ）

148残基目のセリン；
Serine at position 148 （アミノ酸一般名称）
Serine148（アミノ酸一般名称）
Ser-148 （3文字表記）
Ser148 （3文字表記）
S-148 （1文字表記）
S148 （1文字表記）
…

AI(テキストマイニング)による学習

酵素反応を担うアミノ酸（3文字表記と1文字表記）と官能基の例



今年度の進捗①

①酵素反応キュレーション用基本設定

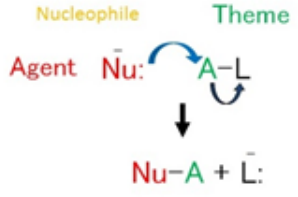
酵素反応キーワード等の分類

- **Entity types**
  - Functional groups
  - Amino acids (AAcid)
  - Cofactors (Cofac)
  - compounds other than cofactors
  - Enzymes
- **Attributes of entities**
  - Catalytic roles
  - Reactive parts in molecules
  - Reaction states
  - Characteristics of entities
- **Reaction events**
  - Reaction step
  - Reaction type
  - Mechanism type of reaction
  - Others
- **Others**

反応事象の定義、設定

Relations: Event type、Agent、Themeの定義

Nucleophilic attack（求核攻撃）(Event type)の例



Agent　　Event type　　Theme

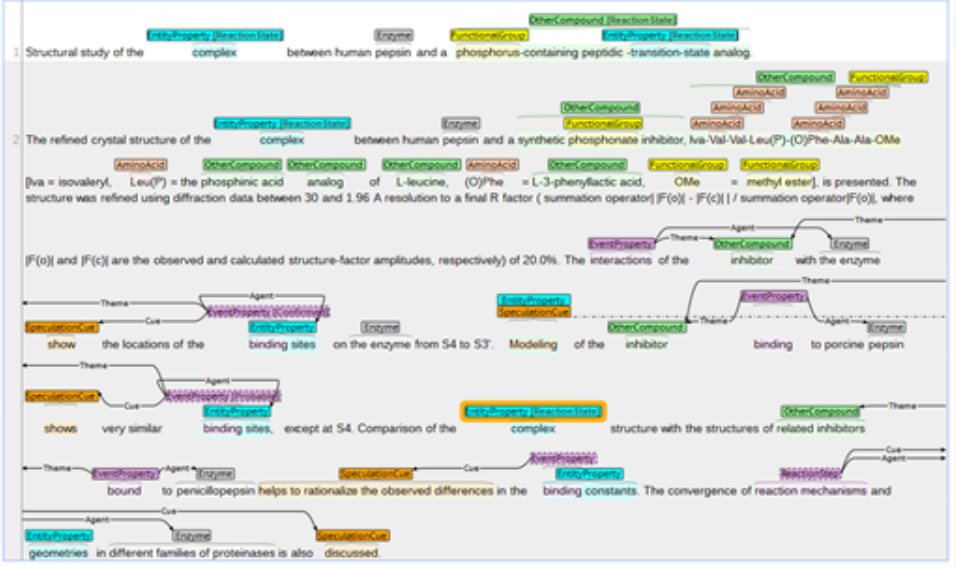
Active-site-Nu acts as a nucleophile to attack on A
= Active-site-Nu makes a nucleophilic attack on A
= (Nucleophilic) active-site-Nu attacks A
=A is attacked by active-site Nu

②EzCatDB中文献の要旨文章のキーワード解析

加水分解酵素、転移酵素のpubmed要旨中に、
主要なキーワードを3種類以上含む要旨等、292件選抜

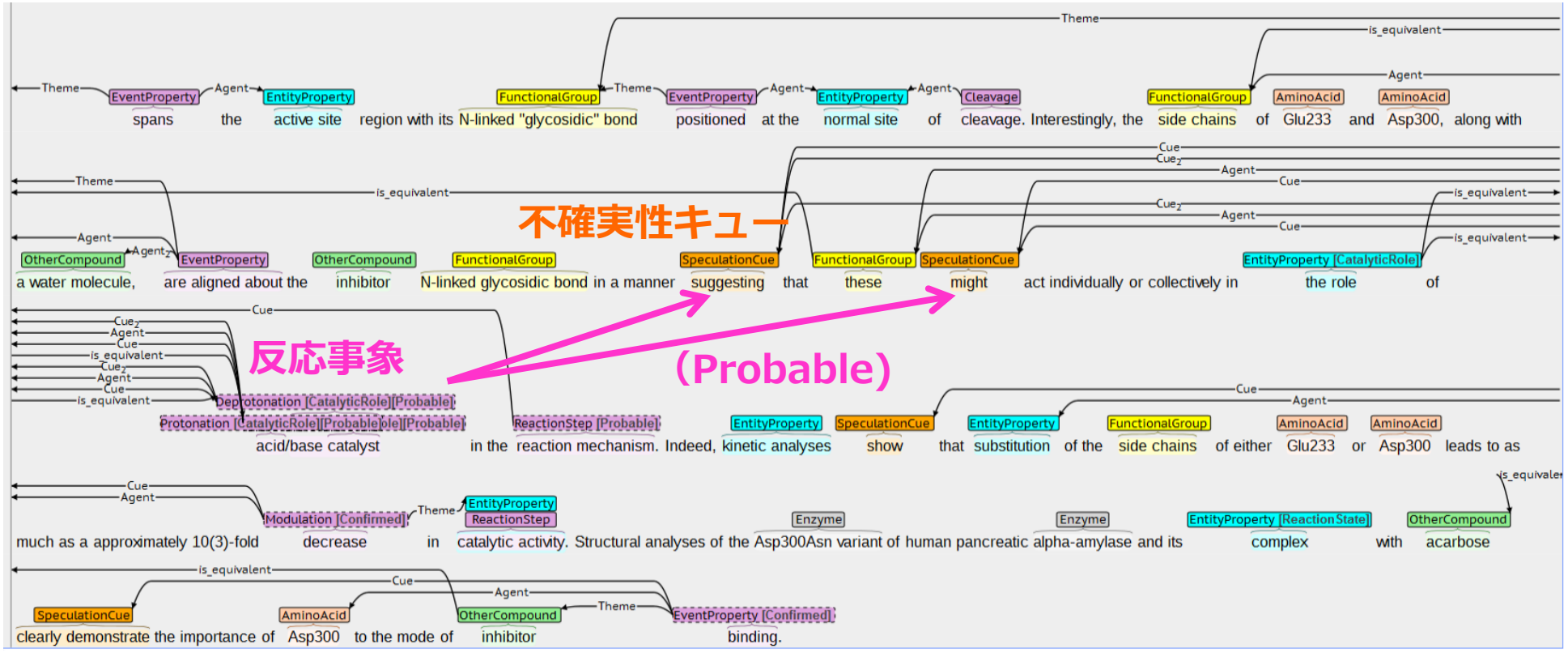
③Pubmed文献要旨文章のキュレーション開始

キュレーションシステム・Bratの設定→キュレーション



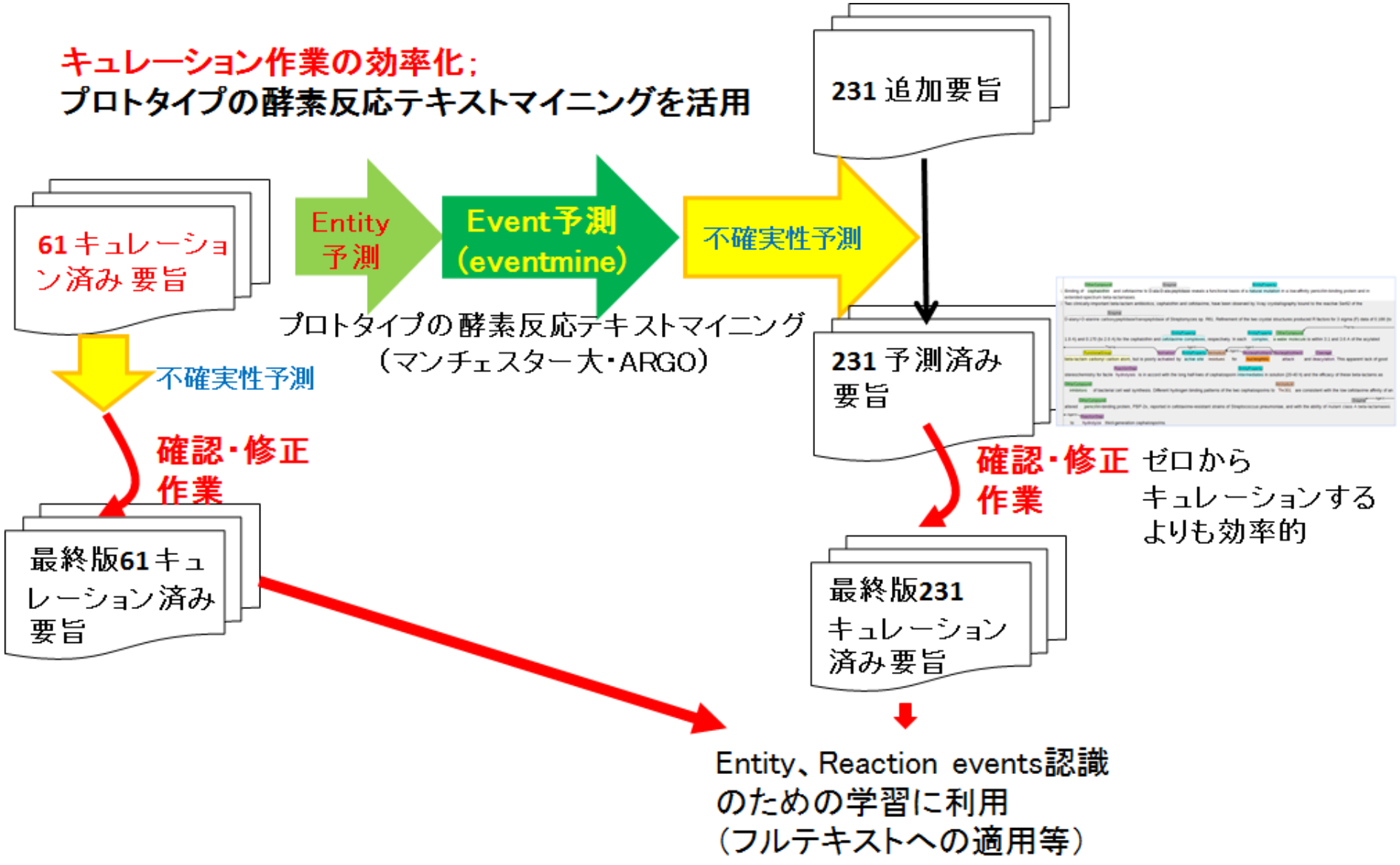
反応事象の不確実性

- 不確実性のレベル
  - Confirmed；実験的に実証済
  - Probable；可能性大
  - Unlikely；可能性小、見込なし
  - Negation；明確な否定形
- キュー（合図）となる表現
  - ”indicate”, “confirm”, etc.
  - ”suggest”, “may”, etc.
  - “unlikely”, etc.
  - ”not”, etc.



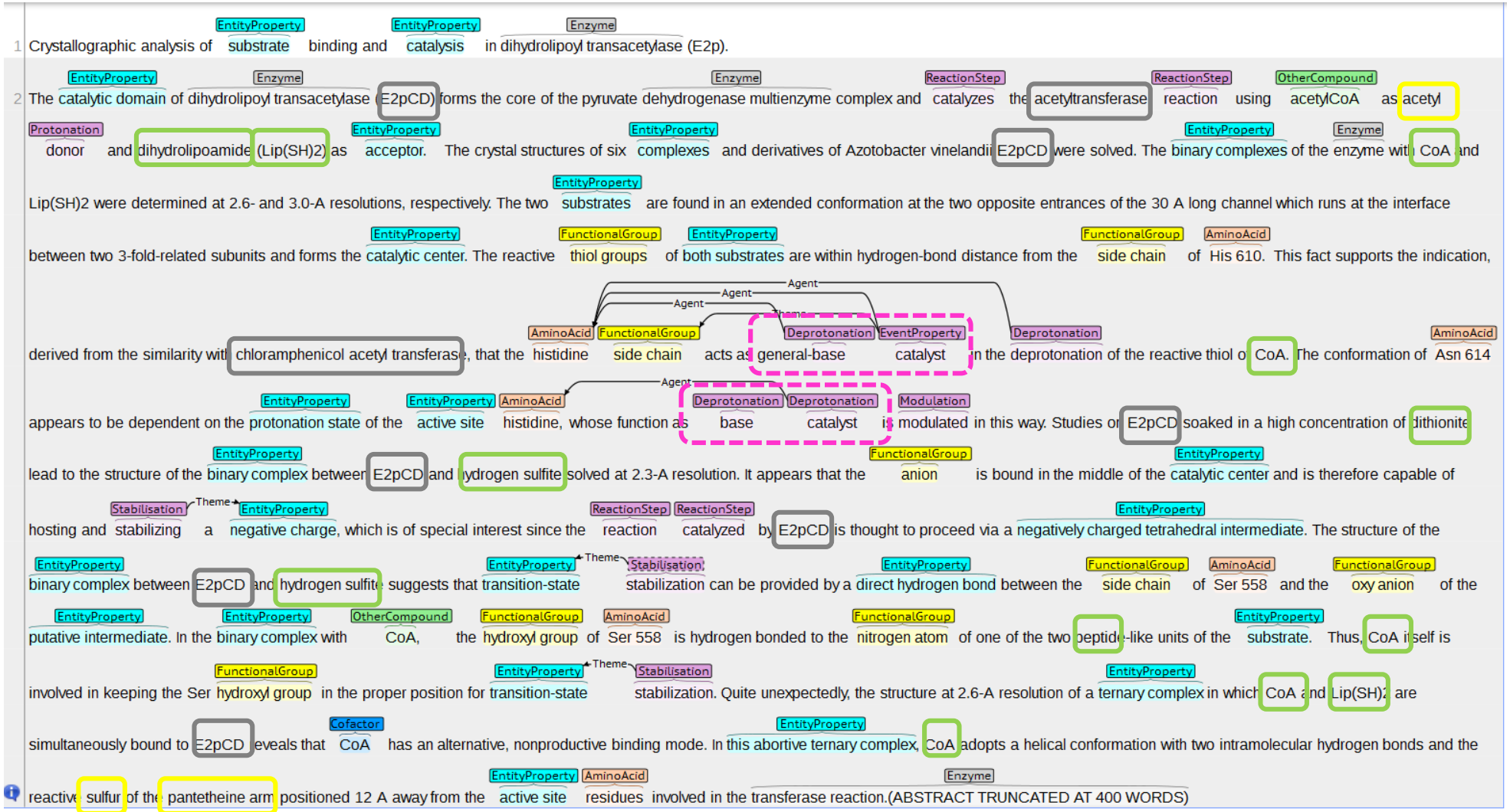
今年度の進捗②

キュレーション作業の効率化；
プロトタイプの酵素反応テキストマイニングを活用



今年度の進捗③

61要旨の学習データによる予備的な予測結果の例



メリット；

- 多くの**反応イベント**を同定可能

問題点；

- 同定できないエンティティ（False negatives）
- 1フレーズが複数のワードに分れて予測される

➡ 改良点

今後の予定

- ARGOによる予測の実施；フルテキストへの適用
  - エンティティ予測
  - 反応イベント予測
  - 不確実性予測
- 学習データの拡充
  - 加水分解酵素、転移酵素以外の酵素へ拡充
    - 異性化酵素、リアーゼ酵素等
    - 新たな反応キーワードの定義
- ARGO以外のアルゴリズムの適用
  - テキスト関係構造認識：キーワードの文中での位置関係の解析