



「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

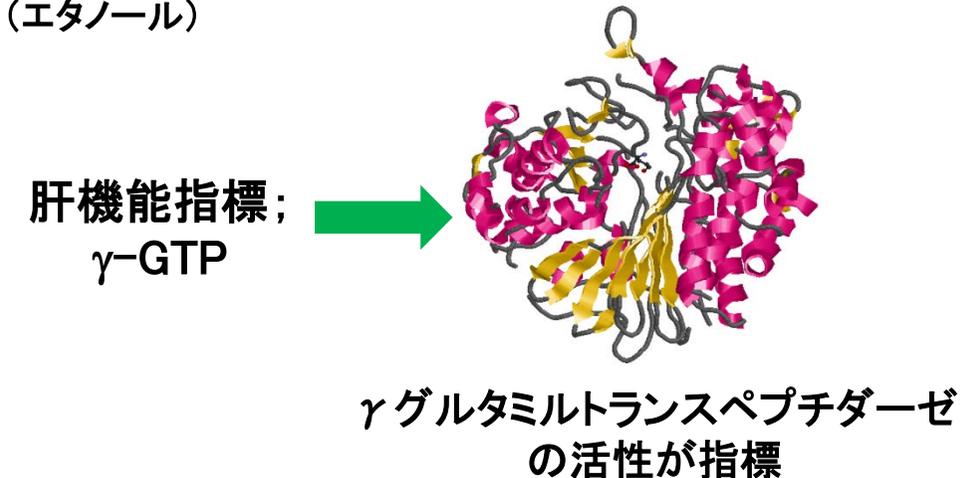
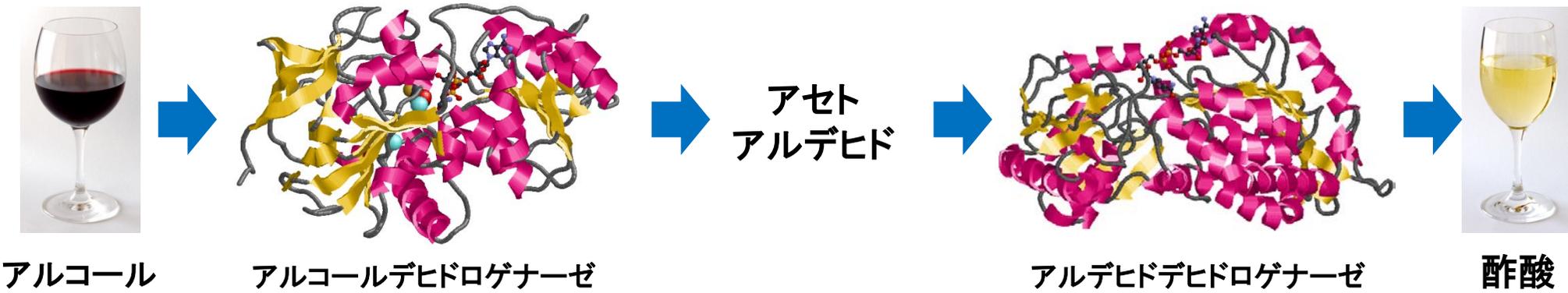
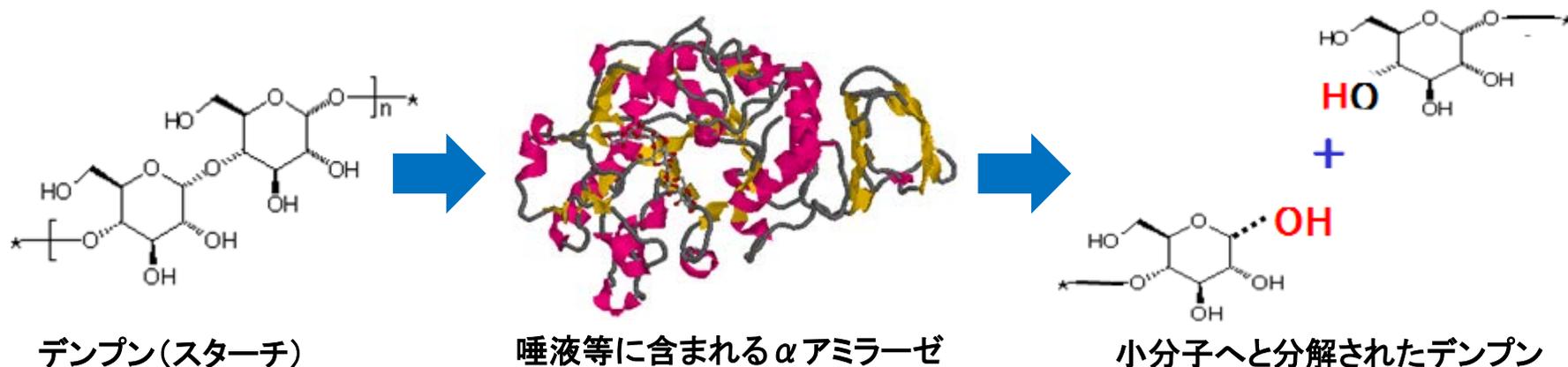
国立研究開発法人 産業技術総合研究所

長野 希美

国立研究開発法人 産業技術総合研究所

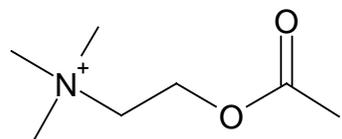
国立研究開発法人 新エネルギー・産業技術総合開発機構

自然界における多様な酵素

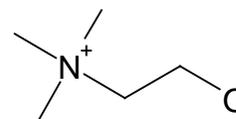


自然界には、5千種類以上の酵素があるとされ、あらゆる反応を担っている。

アセチルコリンエステラーゼ (AChE)

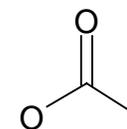


神経伝達物質;
アセチルコリン



コリン

+



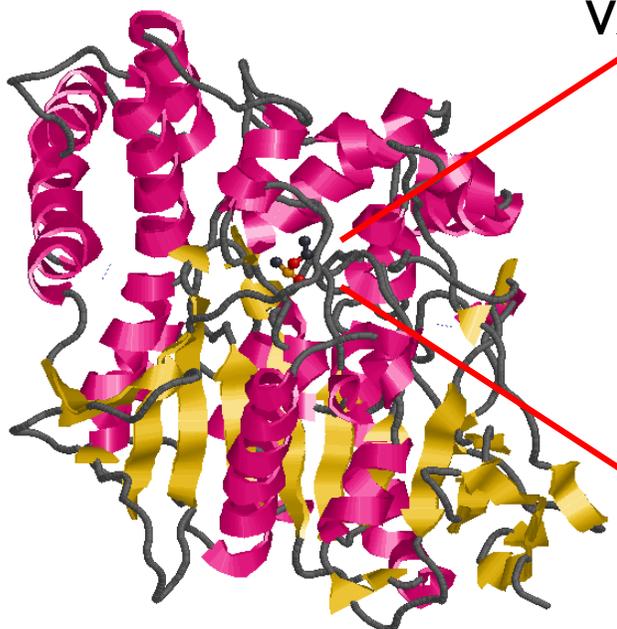
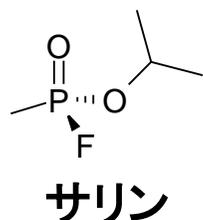
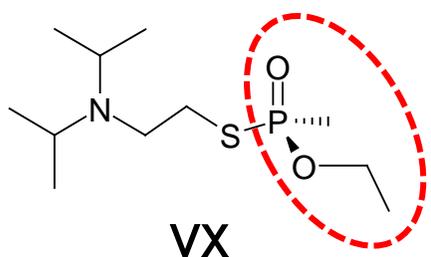
酢酸

アセチルコリンを分解

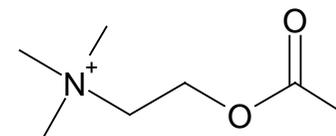
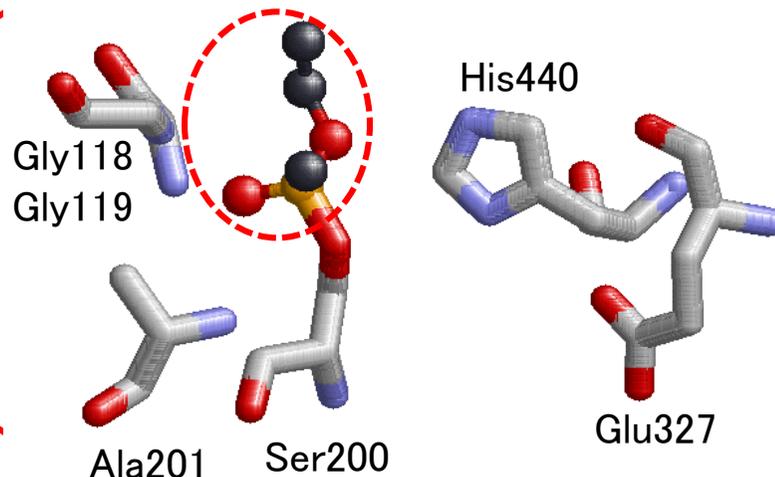
神経ガス

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



**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**

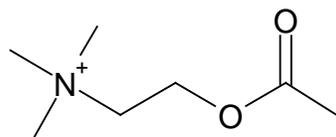
本来、分解すべき、アセチルコリンが結合できず、アセチルコリンを分解できなくなる

● 数々の酵素DB

- IUBMB-Enzyme nomenclature
- IntEnz
- BRENDA
- KEGG
- ExPASy
- ...

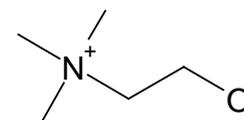
従来の酵素分類の問題点

酵素反応はブラックボックス状態



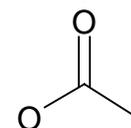
神経伝達物質;
アセチルコリン

AChEの場合



コリン

+



酢酸

基質・産物に基づいた酵素の分類

● 酵素反応データベース

産総研; EzCatDB

酵素エントリ数; 878

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; 335

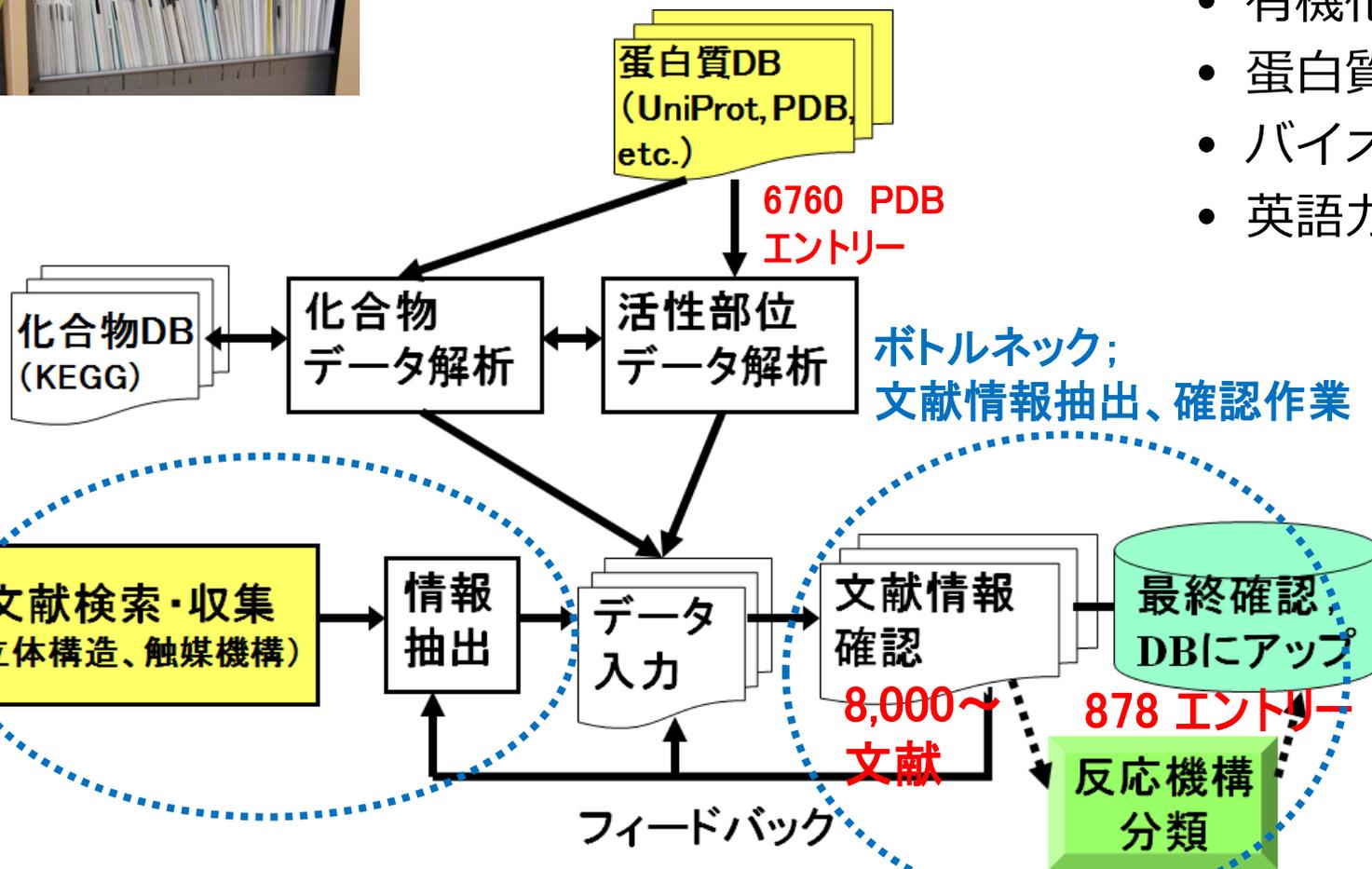
- 酵素反応機構は重要にも関わらず、酵素反応データベースは数少ない。
- 数千の酵素に対して、エントリ数も数百程度と少ない。

• 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



ARCO

英国マンチェスター大学
テキストマイニング・システム

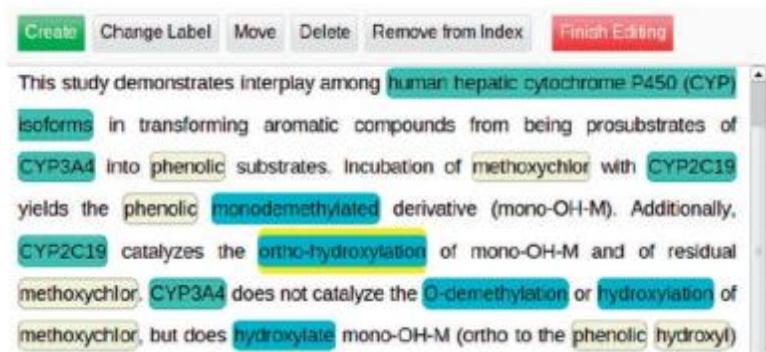
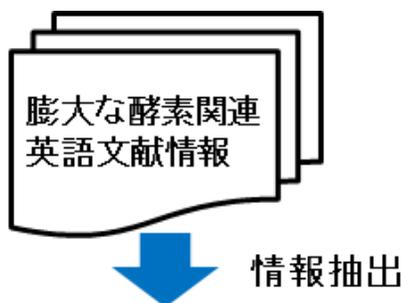
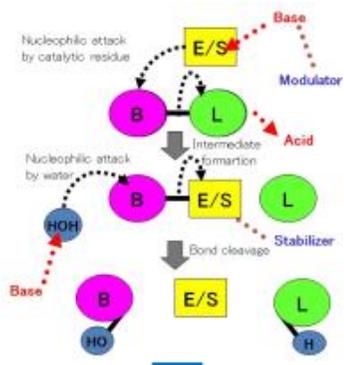


Fig.3 of Rak, *et al.*, (2014) *Database (Oxford)* 2014, pii:bau070.

複雑な酵素反応機構



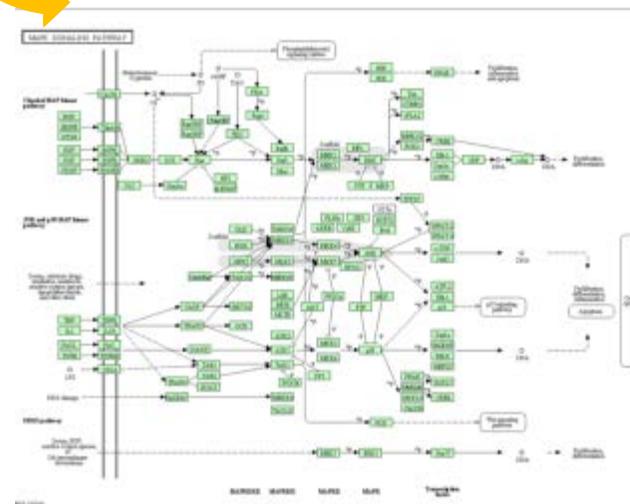
↓ 分類・登録

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

酵素反応データベース・EzCatDB

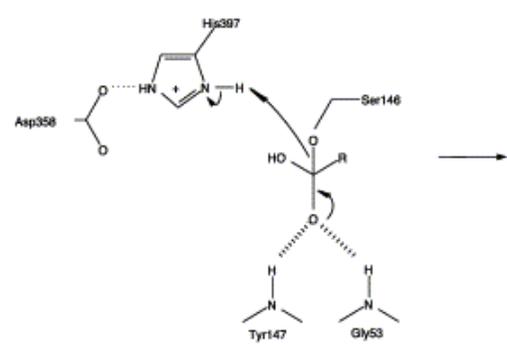
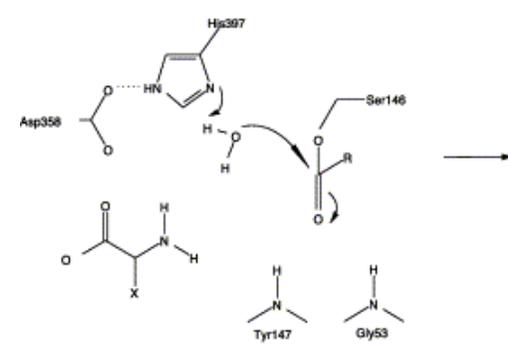
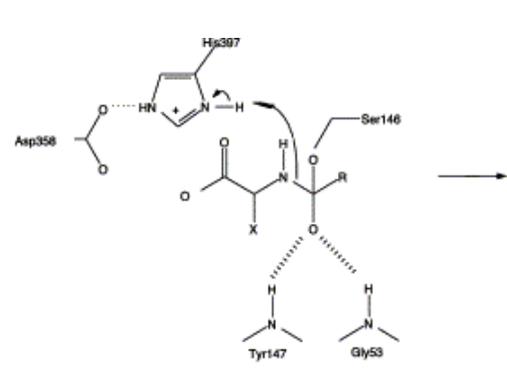
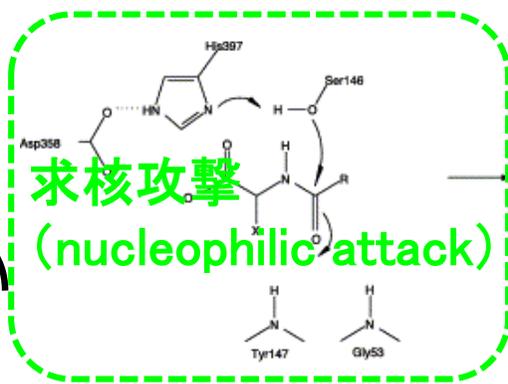


MAPK情報伝達系データの例 @KEGG

セリンプロテアーゼの反応機構図

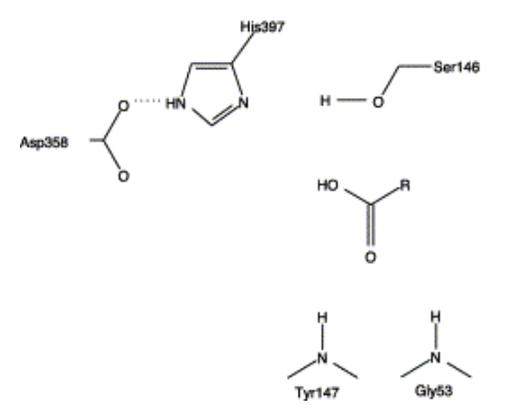
- 文献における反応機構の表現

- 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



①酵素反応キュレーション用基本設定

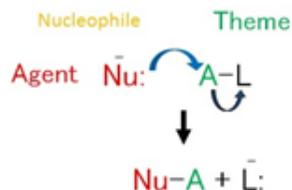
酵素反応キーワード等の分類

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

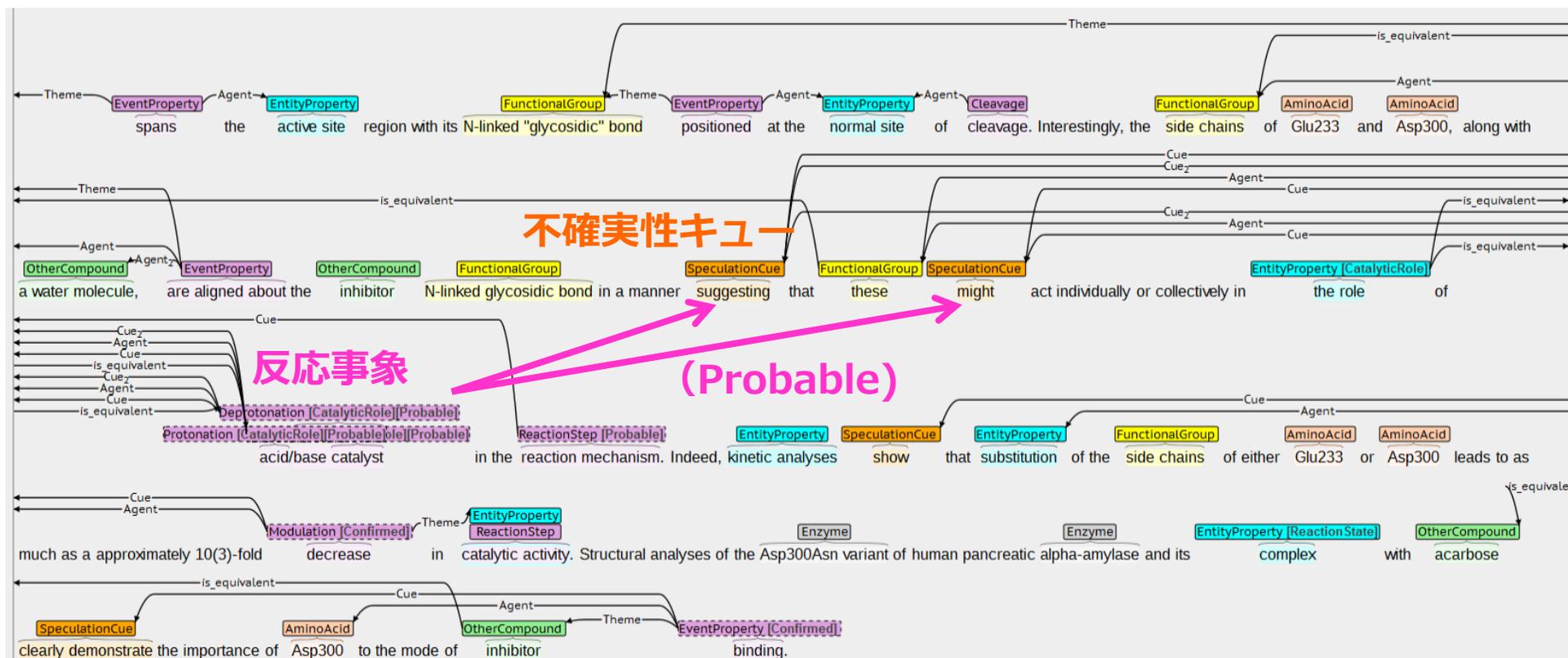


③Pubmed文献要旨文章のキュレーション開始

キュレーションシステム・Bratの設定→キュレーション

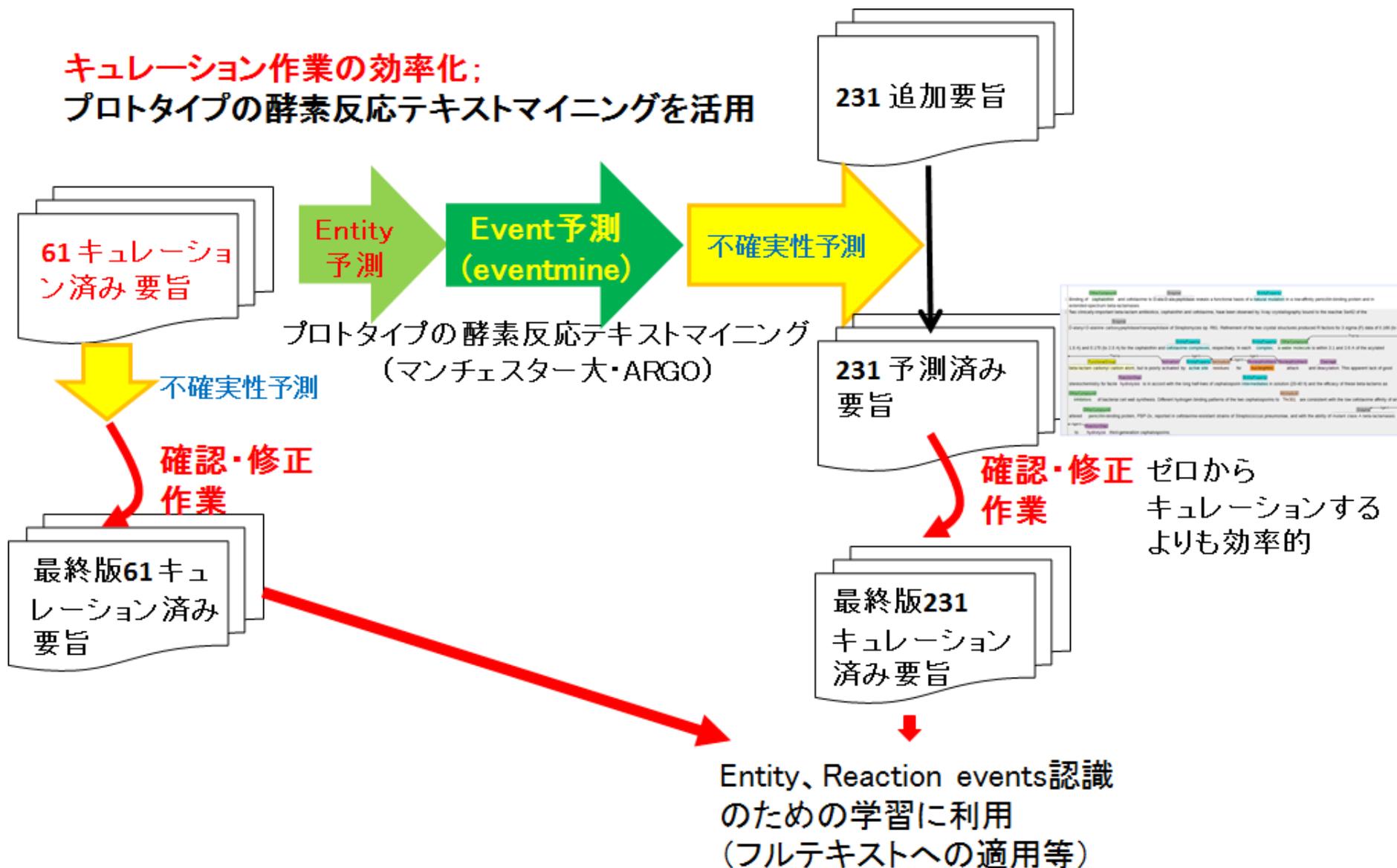
反応事象の不確実性

- 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



61要旨の学習データによる予備的な予測結果の例

1 Crystallographic analysis of substrate binding and catalysis in dihydrolipoyl transacetylase (E2p).

2 The catalytic domain of dihydrolipoyl transacetylase (E2pCD) forms the core of the pyruvate dehydrogenase multienzyme complex and catalyzes the acetyltransferase reaction using acetylCoA as acetyl donor and dihydrolipoamide (Lip(SH)₂) as acceptor. The crystal structures of six complexes and derivatives of Azotobacter vinelandii E2pCD were solved. The binary complexes of the enzyme with CoA and Lip(SH)₂ were determined at 2.6- and 3.0-Å resolutions, respectively. The two substrates are found in an extended conformation at the two opposite entrances of the 30 Å long channel which runs at the interface between two 3-fold-related subunits and forms the catalytic center. The reactive thiol groups of both substrates are within hydrogen-bond distance from the side chain of His 610. This fact supports the indication, derived from the similarity with chloramphenicol acetyl transferase, that the histidine side chain acts as general-base catalyst in the deprotonation of the reactive thiol of CoA. The conformation of Asn 614 appears to be dependent on the protonation state of the active site histidine, whose function as base catalyst is modulated in this way. Studies of E2pCD soaked in a high concentration of dithionite lead to the structure of the binary complex between E2pCD and hydrogen sulfite solved at 2.3-Å resolution. It appears that the anion is bound in the middle of the catalytic center and is therefore capable of hosting and stabilizing a negative charge, which is of special interest since the reaction catalyzed by E2pCD is thought to proceed via a negatively charged tetrahedral intermediate. The structure of the binary complex between E2pCD and hydrogen sulfite suggests that transition-state stabilization can be provided by a direct hydrogen bond between the side chain of Ser 558 and the oxy anion of the putative intermediate. In the binary complex with CoA, the hydroxyl group of Ser 558 is hydrogen bonded to the nitrogen atom of one of the two peptide-like units of the substrate. Thus, CoA is itself involved in keeping the Ser hydroxyl group in the proper position for transition-state stabilization. Quite unexpectedly, the structure at 2.6-Å resolution of a ternary complex in which CoA and Lip(SH)₂ are simultaneously bound to E2pCD reveals that CoA has an alternative, nonproductive binding mode. In this abortive ternary complex, CoA adopts a helical conformation with two intramolecular hydrogen bonds and the reactive sulfur of the pantetheine arm positioned 12 Å away from the active site residues involved in the transferase reaction. (ABSTRACT TRUNCATED AT 400 WORDS)

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