資料6－2

2010年11月2日

# H22年度達成目標と進捗について

## 「専門用語辞書管理システムと専門用語解析技術の開発」

奈良先端科学技術大学院大学

松本裕治

1. 専門用語辞書システムの開発
2. 専門用語解析技術の開発
3. 専門用語抽出ツールの設計と開発

# 今年度の進捗

1. 専門用語辞書システム
   - 仕様設計とMySQLでの実装は完了(Cradle-LSD). 項目の追加削除の柔軟性を考え，MongoDBへ移植中
2. 専門用語解析技術
   - 用語解析システムは昨年度実装. 学習データの蓄積作業は予定通り.
3. バイオ医療専門用語抽出技術
   A) 専門用語を検索するプロトタイプの構築
      i. バイオ医療専門用語の検索
      ii. 類似文脈の検索
   B) コーパスでの出現文脈を利用して専門用語類似度を計算する新しい方法の開発

# A) 専門用語を検索するプロトタイプの構築

トップページ（http://cl.naist.jp/kazuo-h/）:

トップページ　バイオ医療専門用語の検索　類似文脈の検索

## バイオ医療専門用語と類似文脈の検索

コーパスとして PNE (蛋白質・核酸・酵素)を，シソーラスとして LSD (ライフサイエンス辞書)（2008年度版）を用い，次の2つの検索を行います．

- PNEと LSD に対するバイオ医療専門用語の検索（クエリ＝専門用語）
- PNE に対する類似文脈の検索（クエリ＝文脈＋専門用語）

PNE に対する専門用語の検索には，sary を用います．類似文脈の検索では，クエリとして与えられる文脈，および，PNEに登場する専門用語の周辺文脈を，MeCab と CaboCha を用いて抽出し，それら文脈間のコサイン類似度を計算します．

将来的にはシソーラスに未登録の専門用語（新しい治療方法等）の新規登録を支援することが目標であり，より精度の高い類似度計算方法を現在開発中です．

# (i)バイオ医療専門用語の検索

## 検索画面:

トップページ　バイオ医療専門用語の検索　類似文脈の検索

### バイオ医療専門用語の検索(クエリ＝専門用語)

クエリ: 急性骨髄性白血病 [Submit query]

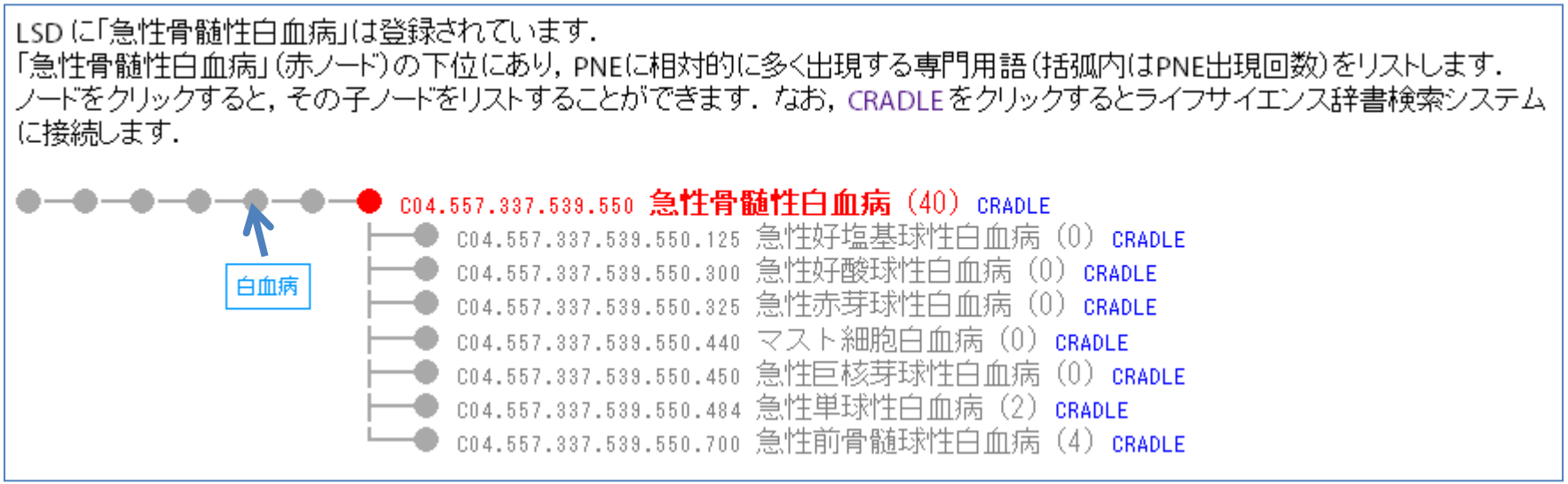
## 検索結果(PNE検索):

「急性骨髄性白血病」を含む文脈は，PNEに10件以上あります.

| | |
|---|---|
| 前者は成熟傾向を示す | **急性骨髄性白血病** (AML) を表現型とする |
| 最近では，X染色体DNAの多型性を指標として | **急性骨髄性白血病**，子宮平滑筋腫，腎に発生するウィルムス腫瘍が単クローン性で |
| たとえば，**急性骨髄性白血病** 細胞では34人の患者中19例にp53蛋白質が増加することが見い | |
| **急性骨髄性白血病** 患者 (24歳，男性) 末梢血 | |
| **急性骨髄性白血病** 株細胞としてML-1以外にKG-1も広く使われている | |
| **急性骨髄性白血病** の腫瘍性芽球も単クローン性である | |
| 先述のとおり，**急性骨髄性白血病** の一部の症例では，赤血球や顆粒球も腫瘍性芽球細胞と同じクロー | |
| これから述べるHL-60細胞は，**急性骨髄性白血病** 患者 (FAB分類M_2) (以前はM_3とされていたが，近 | |
| この後遺症の期間は，同年齢で原爆被曝した人の **急性骨髄性白血病** の発生期間 (図7) とほぼ一致する | |
| 図7は，Ichimaruらが **急性骨髄性白血病** (AML) と急性リンパ性白血病 (ALL) に分けて発表し | |

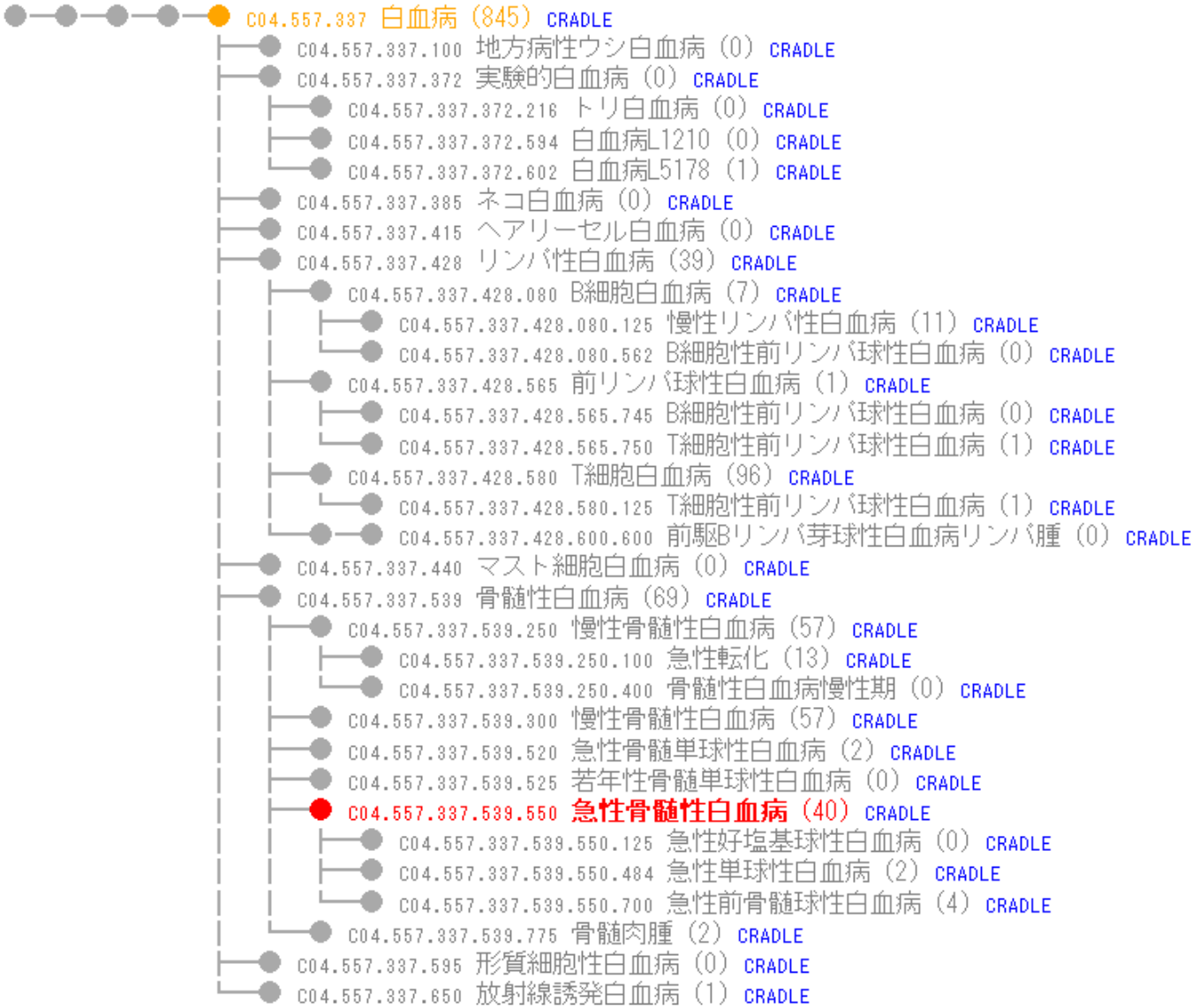
# (i)バイオ医療専門用語の検索

検索結果(LSD検索):

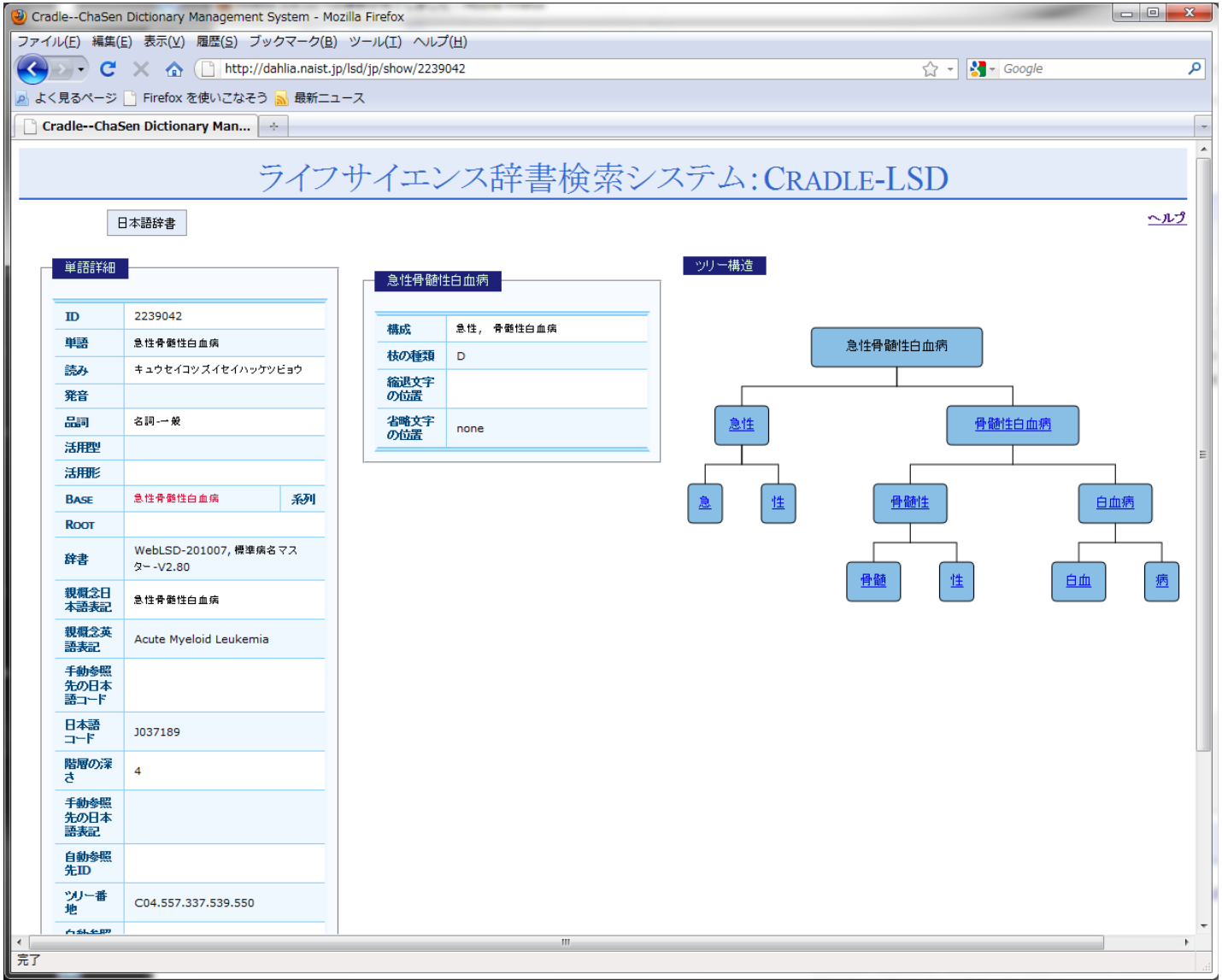



- Cradle-LSD(ライフサイエンス辞書検索システム)への接続も可能.
- ノードをクリックすると, その子ノードをリストする(次のスライド参照).

LSD に「急性骨髄性白血病」は登録されています．
「急性骨髄性白血病」（赤ノード）と「白血病」（橙ノード）の下位にありPNEに相対的に多く出現する専門用語（括弧内はPNE出現回数）と，「白血病」（橙ノード）の子ノードとなる専門用語をリストします．なお，CRADLE をクリックするとライフサイエンス辞書検索システムに接続します．



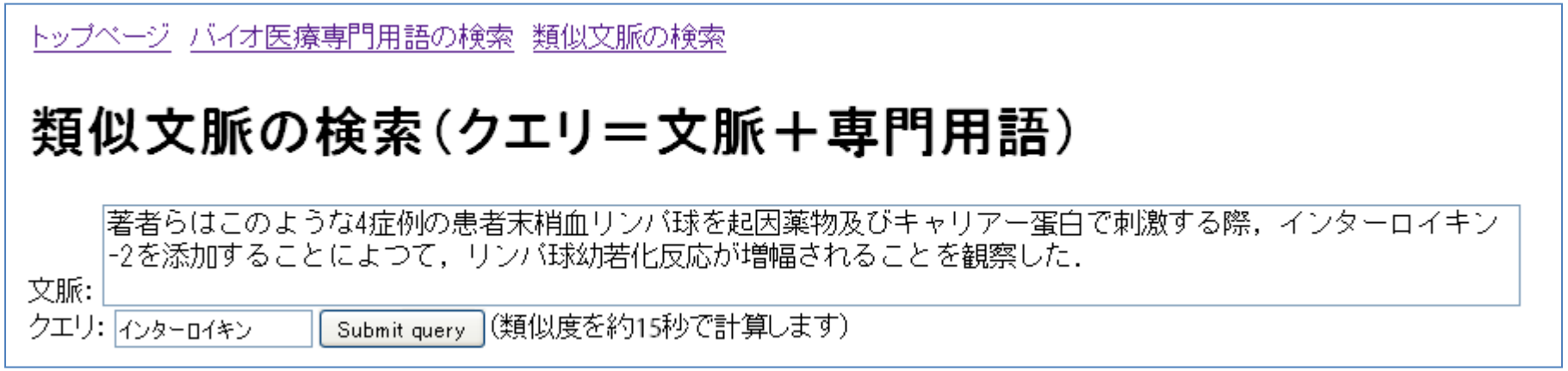
# 専門用語辞書システム (Cradle-LSD)へのリンク



# (ii)類似文脈の検索

目的:クエリとして与えられる文脈，および，PNEに登場する専門用語の周辺文脈を，MeCabとCaboChaを用いて抽出し，それら文脈間のコサイン類似度を計算する.

検索画面:



# (ii)類似文脈の検索

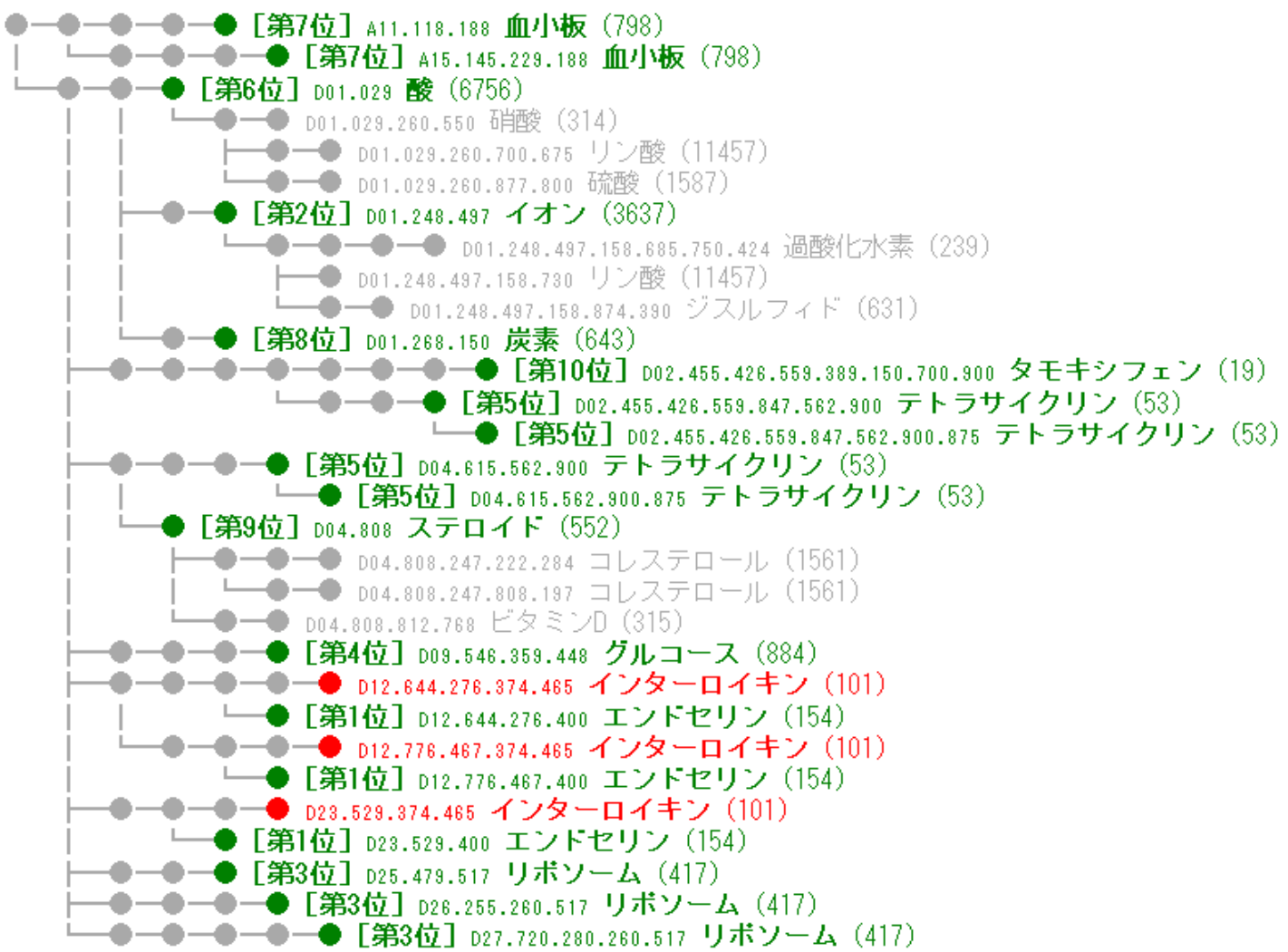
## 検索結果:

「著者らはこのような4症例の患者末梢血リンパ球を起因薬物及びキャリアー蛋白で刺激する際,インターロイキン-2を添加することによつて,リンパ球幼若化反応が増幅されることを観察した.」
に類似する文脈を,PNEから検索した結果(コサイン類似度による10位までのランキング)を示します.

| | | | | |
|---|---|---|---|---|
| 1 | そして外因性の | エンドセリン | -1を添加すると,細胞が収縮した | 0.286 |
| 2 | 実際,ATP,Mgイオン,K^+ | イオン | を複合体の溶液に添加したところ,すぐさま沈殿が生じ, | 0.273 |
| 3 | PE | リボソーム | を添加した細胞を経時的に観察すると,15分後からブリ | 0.266 |
| 4 | | グルコース | を添加することによって,欠損株の表現型は抑制される | 0.265 |
| 5 | しかし, | テトラサイクリン | を添加することで,TetRがTetO配列に結合するこ | 0.263 |
| 6 | そして,この胚様体形成時にレチノイン | 酸 | を添加することによって神経系の分化がさらに促進される | 0.263 |
| 7 | | 血小板 | にスフィンゴシンキナーゼの阻害剤を添加すると,S1P | 0.259 |
| 8 | ゲナーゼは哺乳動物では5位,12位あるいは15位の | 炭素 | 原子に1分子の酸素を添加するものが知られており,それ | 0.255 |
| 9 | しかし,ステロイドホルモン応答細胞に | ステロイド | ホルモンを添加したり,幼若ラットにテストステロン,エ | 0.255 |
| 10 | この際, | タモキシフェン | を添加しておくとエストロゲン作用は打ち消され,無添加 | 0.252 |

# クエリおよび，クエリと類似する専門用語のLSDにおける位置

上表に示した類似文脈に関わる専門用語（緑ノード）と「インターロイキン」（赤ノード）に対して，LSDにおいて下位にあり，PNEに相対的に多く出現する専門用語をリストします．



# プロトタイプの構築に関する補足

- 今後，シソーラスに未登録の専門用語（新しい治療方法等）の新規登録を支援することを目指す．
- 現在は，より精度の高い類似度計算方法を開発している（次のスライド以降で説明）．

# 2. 専門用語類似度計算方法の開発

- 標準的な方法は，出現文脈を周辺単語をbag-of-words とみなして類似度を測る.
  - たとえば，コサイン類似度.
- 開発している手法は，鹿島らのグラフカーネル [Kashima et al., ICML 2003] を応用し，専門用語が出現する文の係り受け木の上でランダムウォークを行うことにより，周辺単語の構造を捉えてそれらの類似度を測る.

# 標準的な方法

- 出現文脈を周辺単語をbag-of-words とみなして類似度を測る.
  - 次の例 :

    **例文1**: Androgen receptor blockade might be potentiated by a reduction of serum **androgens** .

    **例文2**: These results can not be explained by receptor down-regulation due to higher levels of **mineralocorticoids** in PIH .

    を用い,

    androgens と mineralocorticoids のコサイン類似度計算を, 次のスライドで示す.

**androgens** の特徴ベクトル

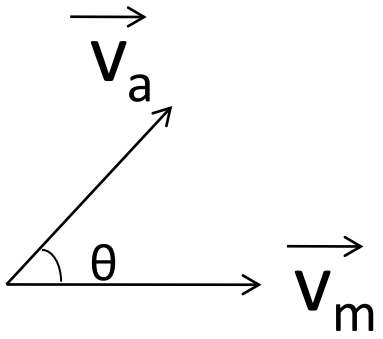
$$\overrightarrow{V_a} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

**mineralocorticoids** の特徴ベクトル

$$\overrightarrow{V_m} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

| | |
|---|---|
| 1 | receptor |
| 1 | blockade |
| 1 | might |
| 1 | be |
| 1 | potentiated |
| 1 | by |
| 1 | a |
| 1 | reduction |
| 1 | of |
| 1 | serum |
| 0 | explained |
| 0 | down-regulation |
| 0 | due |
| 0 | to |
| 0 | higher |
| 0 | levels |
| 0 | in |
| 0 | PIH |

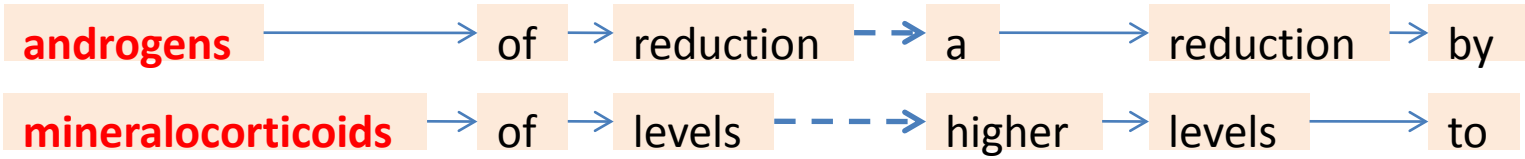
コサイン類似度： 2つの単語の類似度を，
2つの単語に対応する特徴ベクトルが
成す角度θのコサインと定義する．

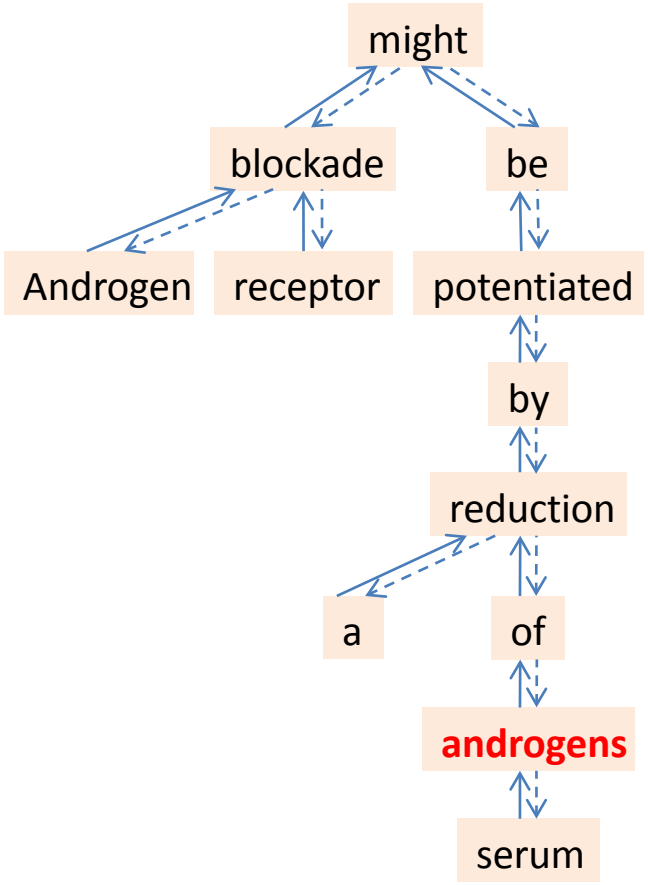


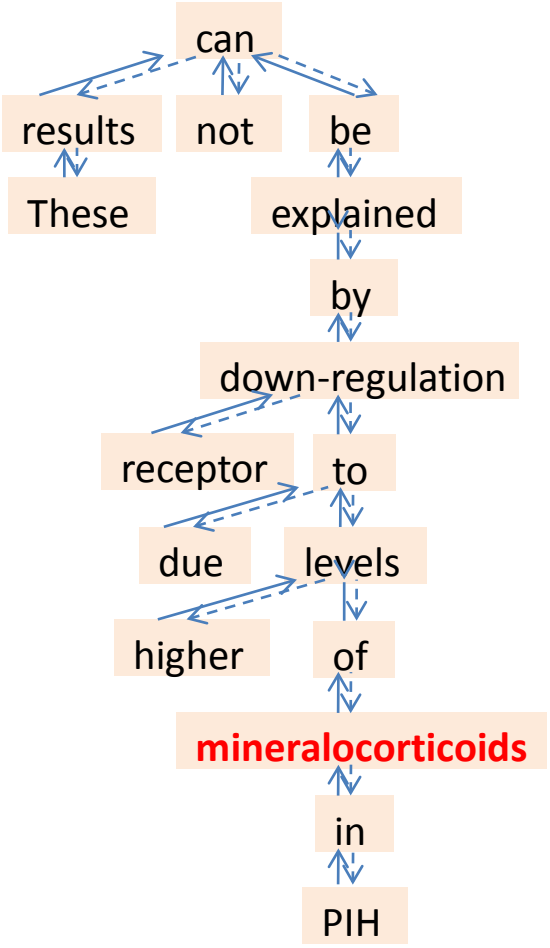
# 標準的な方法の短所

- 例文中で “a reduction of androgens”, “higher levels of mineralocorticoids” という句を構成して出現しているが，こうした句の構造を，周辺単語をbag-of-words として扱う手法では捉えることができない.
  - androgens と mineralocorticoids は，ホルモンの一種であり，分泌量の増減，あるいは，体液中の濃度を記述する文脈で用いられることが多い.

# 係り受け木の上で並行ランダムウォークを行う利点：周辺文脈の構造を捉えることができること.



# シソーラスマッピングタスクによる提案手法の評価

- 結果（次のスライドの表）の見方
  - クエリと親子または兄弟の関係にある最上位専門用語のランクの平均値を示す.
  - N はランダムウォークの長さ，$\gamma$ はランダムウォークの停止確率である.
  - N= 1は，係り受け木の構造情報を用いないコサイン類似度による手法の結果に相当する.

- 提案手法はN ≤ 8，$\gamma$ = 0.001 のときにベストの平均ランク 37.21 位を得て，コサイン類似度による 42.48 位を約 5 ポイント上回る結果となった.

# シソーラスマッピングタスクでの提案手法の評価実験の結果

| $\gamma$ | 0.001 | 0.005 | 0.01 | 0.05 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 |
|---|---|---|---|---|---|---|---|---|---|
| N＝1 | 42.48 | 42.48 | 42.48 | 42.48 | 42.48 | 42.48 | 42.48 | 42.48 | 42.48 |
| N≦2 | 40.43 | 40.43 | 40.43 | 40.31 | 39.70 | **38.99** | **39.54** | **40.68** | **41.29** |
| N≦3 | 38.49 | 38.5 | 38.50 | **38.21** | **38.85** | 40.72 | 41.36 | 41.82 | 42.06 |
| N≦4 | 39.17 | 39.12 | 38.88 | 38.34 | 39.92 | 41.17 | 41.68 | 41.91 | 42.07 |
| N≦5 | 38.08 | 37.79 | **37.81** | 39.95 | 40.92 | 41.47 | 41.75 | 41.91 | 42.06 |
| N≦6 | 37.94 | **37.56** | 38.18 | 40.40 | 41.14 | 41.51 | 41.76 | 41.91 | 42.06 |
| N≦7 | 37.67 | 38.3 | 39.31 | 40.92 | 41.24 | 41.53 | 41.76 | 41.91 | 42.06 |
| N≦8 | **37.21** | 38.73 | 39.94 | 41.03 | 41.26 | 41.53 | 41.76 | 41.91 | 42.06 |
| N≦9 | 38.26 | 39.7 | 40.17 | 41.16 | 41.30 | 41.53 | 41.76 | 41.91 | 42.06 |
| N≦10 | 38.9 | 40.1 | 40.45 | 41.19 | 41.30 | 41.53 | 41.76 | 41.91 | 42.06 |