厚生労働科学研究費補助金
(医療安全・医療技術評価総合研究事業)

テキストマイニングによる薬物有害事象の自動抽出を目的とした オントロジー構築とシステム開発に関する研究

平成19年度　総括研究報告書

Annual Report
Grant-in-Aid for Research on Medical Safety and Medical Technology Evaluation
Supported by the Ministry of Health, Labor and Welfare, Japan in 2007
(Chief Researcher: Shuji KANEKO, Ph.D.)

平成20年(2008)年3月

主任研究者　金子　周司

目　　次

厚生労働科学研究費補助金（医療安全・医療技術評価総合研究事業）
研究報告書

テキストマイニングによる薬物有害事象の自動抽出を目的としたオントロジー構築とシステム開発

主任研究者　金子周司　京都大学大学院薬学研究科生体機能解析学分野

研究協力者：大武 博（京都府立医科大学医学研究科），藤田信之（製品開発技術評価機構）

[研究要旨]

ライフサイエンス辞書（LSD）は，薬学の諸領域を含む広範な生命科学の論文や教科書で用いられる英語（8万語）および日本語（9万語）の専門用語を，用語の出現頻度を重要性の目安として選出した対訳辞書である。本研究では，LSDに収録されている名詞のうち，病名/症候名，薬物/生体内分子名，解剖/発生部位名，生物名，方法や研究技術を意味する約4万語の対訳について，用語の同義性や上下関係を整理し，既存の専門用語シソーラスであるMeSHとリレーショナルデータベースで関連づけた。また，公開されている病名分類（ICD-10準拠標準病名マスターおよびICH国際医薬用語集MedDRA/J）や薬効分類（WHO ATCおよびMeSH Pharmacological Actions）へのリンクを設けることでLSDに意味と拡張性を付与し，新しい用語を補充した。この結果，約16万語の専門用語をツリー状に整理した2万語の統制語に割り当てたLSDシソーラスをほぼ完成させた。次にこれを応用するため，専門領域の英語テキストにシソーラス収録の日本語対訳タグを付与し，WWWブラウザで視認性良く閲覧できるようにするPerlスクリプトを開発した。さらに，大量の英文テキスト中での用語の共起関係を数値化することによって共起キーワードを収集した。このLSDシソーラスを利用したテキスト処理は，医療文書からの有害事象の検出をはじめとして関係抽出に応用できる優れた手法になると考えられる。

A. 研究目的

本研究は，ゲノム科学における情報科学的手法として発展・応用されつつあるテキストマイニング技術を医薬品の副作用（有害事象）のレポートや医療情報の解析に最適化し，日本語と英語を網羅する医療関連の用語オントロジーをテキスト解析エンジンに実装して，その評価を行いつつ，実効性のある情報解析システムを開発することによって，情報電子化時代を迎える医療における効率良く確かな安全体制の実現を，情報技術的に支援することを目的とした。

研究2年目にあたる19年度は，テキストマインイングを行うために必要な用語について，同義語を網羅し，かつ概念の上下関係が記述された頑強なシソーラスを実用レベルへ拡張させることを目標とした。また，構築した辞書を用いたテキストマイニングエンジンの試作，評価を行い，情報抽出に向けての具体的ステップを実現した。

B. 研究方法

1．概念の構造化によるシソーラス構築

ライフサイエンス辞書（LSD）[1-2]には，生命科学の学術論文で用いられる専門用語が英語，日本語の見出し語がそれぞれ8万ないし9万語収録されており，さらにそれらの対訳関係が 11 万対で規定されている。このうち，病名および症候名，薬物および生体内分子名，解剖および発生部位名，生物名，方法や研究技術を意味する約5万語の対訳について，NLM が定期的に更新している MeSH ツリーの2008年版(2007年11月公開)を用いて，英語によるマッチングを行った。その際，LSD に収録されている単数形の自然テキスト順の用語と，MeSH に多く見られる複数形および階層性を残した語順の表記（例：leukemia, acute など）を一致させるため，MeSH 用語を標準テキスト順に戻し，かつ複数形を単数形に揃える Perl スクリプトで前処理を施し，照合数の改善を行った。

次に，解剖部位名，生物名，方法と技術について MeSH に準拠した階層化を行い，さらに LSD 収録語を各グループに帰属させる作業を行った。また，薬物名を商品名などの通称に対応させるとともに，薬効カテゴリーによるグループ化を行うため，LSD 対訳を WHO ATC 分類および MeSH Pharmacological Actions に関連づけ，各薬物に薬効タグを付与するとともに，新たな薬物名を収録した。また，複数存在する病名の階層化に対応するため，既存の病名分類である ICD-10 準拠標準病名マスターおよびICH 国際医薬用語集 MedDRA/J と関連づけ，さらに LSD に収録されている様々な表記法を可能な限りそれらの分類に帰属させた。

このようにして作成したシソーラスから，関係抽出のための辞書を試作した。辞書構造は，複数存在する英語シノニムの１つ１つを代表的な日本語表記に置換し，さらに病名，物質名など５種類の属性を付与できるよう，１対１のテキストとした。

2．テキストタグ付けスクリプトの制作

PubMed 抄録を題材にして，様々な英語表記を代表的な日本語訳に逐語訳するため，Perl スクリプトを制作した。また，１文中に共起する日本語訳（キーワード）を集計する Perl スクリプトを設計した。処理は次の３段階で行った。

（1）抄録の抽出とセンテンスへの切り分け

（2）英語から日本語への逐語訳[3)]

（3）共起キーワード計数

逐語訳においては，処理されたテキストを WWW ブラウザでカラー表示するため，XML タグを付与する形とした。

それぞれのスクリプトは試行を繰り返し，処理速度の最適化をはかった。

C. 研究結果

1．シソーラスの構築

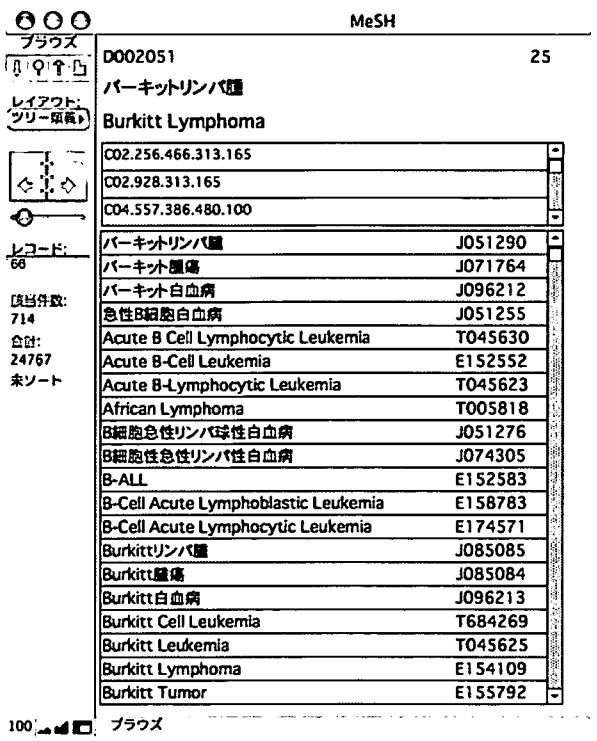
今年度はライフサイエンス辞書（LSD）に収録された 11 万語の専門対訳レコードのうち，有害事象の関係抽出に必要となると考えられる病名および症候名，薬物および生体内分子名，解剖および発生部位名，生物名，方法や研究技術を意味する約5万語の対訳について，用語の同義性や上下関係を整理し，さらに既存の専門用語シソーラスである MeSH 2008 年版とリレーショナルデータベースで動的に関連づけた。また，公開されている病名分類（ICD-10 準拠標準病名マスターおよび ICH 国際医薬用語集 MedDRA/J）や薬効分類（WHO anatomical therapeutic chemical 分類および MeSH Pharmacological Actions）へのリレーションを設けることで LSD に拡張性を付与した。この結果，約 16 万語の専門用語を約 2 万語のツリー状に整理した見出し語（統制語）に割り当てた LSD シソーラスが全体計画の 90%まで構築した。特に，病名と物質名のシノニムテーブルが完成したことで，医療テキストの解析に十分なシソーラスが構築できたものと思われる。

図１－３には，制作したシソーラスから，病名ツリーの一部分，病名シノニム，医薬品シノニムと薬効分類の例を示している。

C09.218.458	聴覚障害 Hearing Disorder
C09.218.458.341	難聴 Hearing Loss
C09.218.458.341.186	聴覚消失 Deafness
C09.218.458.341.186.5	盲聾障害 Deaf-Blind Disorder
C09.218.458.341.186.5	アッシャー症候群 Usher Syndrome
C09.218.458.341.186.5	ウルフラム症候群 Wolfram Syndrome
C09.218.458.341.374	両側性難聴 Bilateral Hearing Loss
C09.218.458.341.562	伝音難聴 Conductive Hearing Loss
C09.218.458.341.750	機能性難聴 Functional Hearing Loss
C09.218.458.341.812	高周波聴力消失 High-Frequency Hearing Loss
C09.218.458.341.849	混合性難聴 Mixed Conductive-Sensorineural Hearing Loss
C09.218.458.341.887	感音難聴 Sensorineural Hearing Loss
C09.218.458.341.887.4	中枢性難聴 Central Hearing Loss
C09.218.458.341.887.4	騒音性難聴 Noise-Induced Hearing Loss
C09.218.458.341.887.7	老人性難聴 Presbycusis
C09.218.458.341.887.8	アッシャー症候群 Usher Syndrome
C09.218.458.341.900	突発性難聴 Sudden Hearing Loss
C09.218.458.341.950	片側性難聴 Unilateral Hearing Loss
C09.218.458.505	聴覚過敏 Hyperacusis
C09.218.458.670	耳鳴 Tinnitus
C09.218.513	耳帯状疱疹 Herpes Zoster Oticus
C09.218.568	内耳疾患 Labyrinth Disease
C09.218.568.120	蝸牛疾患 Cochlear Disease
C09.218.568.217	内リンパ水腫 Endolymphatic Hydrops
C09.218.568.217.500	メニエール病 Meniere Disease
C09.218.568.558	内耳炎 Labyrinthitis
C09.218.568.900	前庭疾患 Vestibular Disease
C09.218.568.900.883	回転性めまい Vertigo
C09.218.705	耳炎 Otitis
C09.218.705.371	内耳炎 Labyrinthitis
C09.218.705.496	外耳炎 Otitis Externa
C09.218.705.663	中耳炎 Otitis Media
C09.218.705.663.652	乳様突起炎 Mastoiditis
C09.218.705.663.680	化膿性中耳炎 Suppurative Otitis Media
C09.218.705.663.683	滲出性中耳炎 Otitis Media with Effusion

図１　病名についての概念ツリー構造

階層の深さはインデントとして表されている。



MeSH

ブラウズ

D002051　25

バーキットリンパ腫

Burkitt Lymphoma

C02.256.466.313.165	
C02.928.313.165	
C04.557.386.480.100	

レコード: 66

該当件数: 714

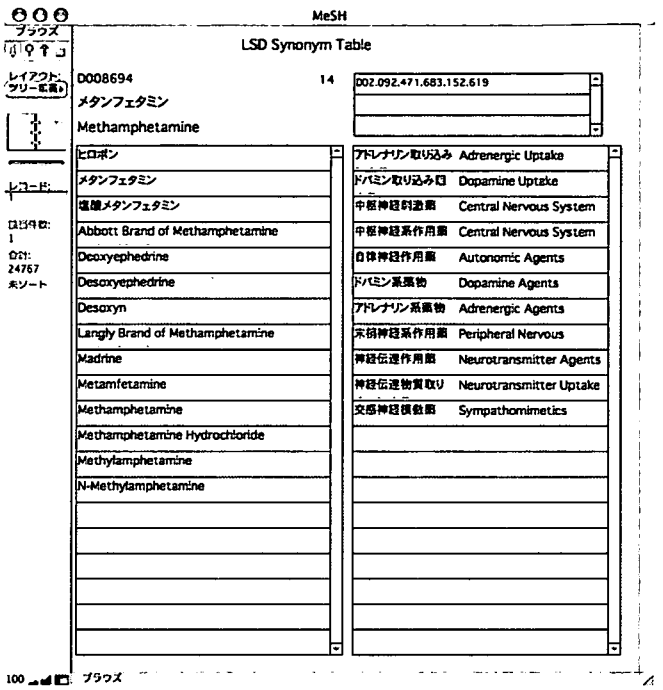
合計: 24767

未ソート

バーキットリンパ腫	J051290
バーキット腫瘍	J071764
バーキット白血病	J096212
急性B細胞白血病	J051255
Acute B Cell Lymphocytic Leukemia	T045630
Acute B-Cell Leukemia	E152552
Acute B-Lymphocytic Leukemia	T045623
African Lymphoma	T005818
B細胞急性リンパ球性白血病	J051276
B細胞性急性リンパ性白血病	J074305
B-ALL	E152583
B-Cell Acute Lymphoblastic Leukemia	E158783
B-Cell Acute Lymphocytic Leukemia	E174571
Burkittリンパ腫	J085085
Burkitt腫瘍	J085084
Burkitt白血病	J096213
Burkitt Cell Leukemia	T684269
Burkitt Leukemia	T045625
Burkitt Lymphoma	E154109
Burkitt Tumor	E155792

100 ブラウズ

図２　病名シノニムテーブルの例



MeSH

ブラウズ

LSD Synonym Table

D008694　14

メタンフェタミン

Methamphetamine

D02.092.471.683.152.619

ヒロポン	アドレナリン取り込み Adrenergic Uptake
メタンフェタミン	ドパミン取り込み D Dopamine Uptake
塩酸メタンフェタミン	中枢神経刺激薬 Central Nervous System
Abbott Brand of Methamphetamine	中枢神経系作用薬 Central Nervous System
Deoxyephedrine	自律神経作用薬 Autonomic Agents
Desoxyephedrine	ドパミン系薬物 Dopamine Agents
Desoxyn	アドレナリン系薬物 Adrenergic Agents
Langly Brand of Methamphetamine	末梢神経系作用薬 Peripheral Nervous
Madrine	神経伝達作用薬 Neurotransmitter Agents
Metamfetamine	神経伝達物質取り Neurotransmitter Uptake
Methamphetamine	交感神経模倣薬 Sympathomimetics
Methamphetamine Hydrochloride	
Methylamphetamine	
N-Methylamphetamine	

レコード:

日該当件数: 1

合計: 24767

未ソート

100 ブラウズ

図３　医薬品シノニムと薬効分類の例

左カラムはシノニムを，右カラムは薬効分類を表す。

なお，これらのデータについては，資料ページにおいて以下のように抜粋データを掲載する。

資料編１（T1～T170 ページ）シソーラスツリー

解剖，生物名，病名，物質名を表す 17,888 語の日英併記の統制語について，最大 10 層のツリーに割り当てた。1 つの統制語が複数のツリー下に属するため，全体でツリー項目数は 36,185 である。

資料編２（D1～D303 ページ）病名シノニム

病名および症候名を表す日英 26,371 語を 4,268 語の統制語へ割り当てた。

資料編３（C1～C428 ページ）

医薬品シノニムおよび薬効分類

生物活性を有する 3,482 種類の化合物について生物活性を表すタグおよび日英異表記 39,105 語を収録した。

2．テキストへのタグ付けスクリプトの開発

専門領域の英語テキストにシソーラス収録の日本語対訳タグを高速に付与するPerlスクリプトを開発した。このタグ付けされたXMLファイルをWWWブラウザでカラー表示する（図４）ことによって，視認性良く英文テキストを閲覧できるようにした。さらに，医薬品と適応症（あるいは有害事象）のように関連する事物やイベントが１文中で共起した場合にそれらを自動検出するPerlスクリプトを開発した（図５）。

これらのスクリプトを連続して用いると，例えば世界標準として用いられている薬理学の教科書Goodman & Gilman's The Pharmacological Basis of Therapeutics 10th editionの全テキスト（約8 MB，100万ワード超）に対して，市販されている普通のパソコンを使ってもわずか70秒ですべての医薬品と適応症や副作用などにタグ付けを行い，それらの出現頻度や共起関係を計数することが可能となった。

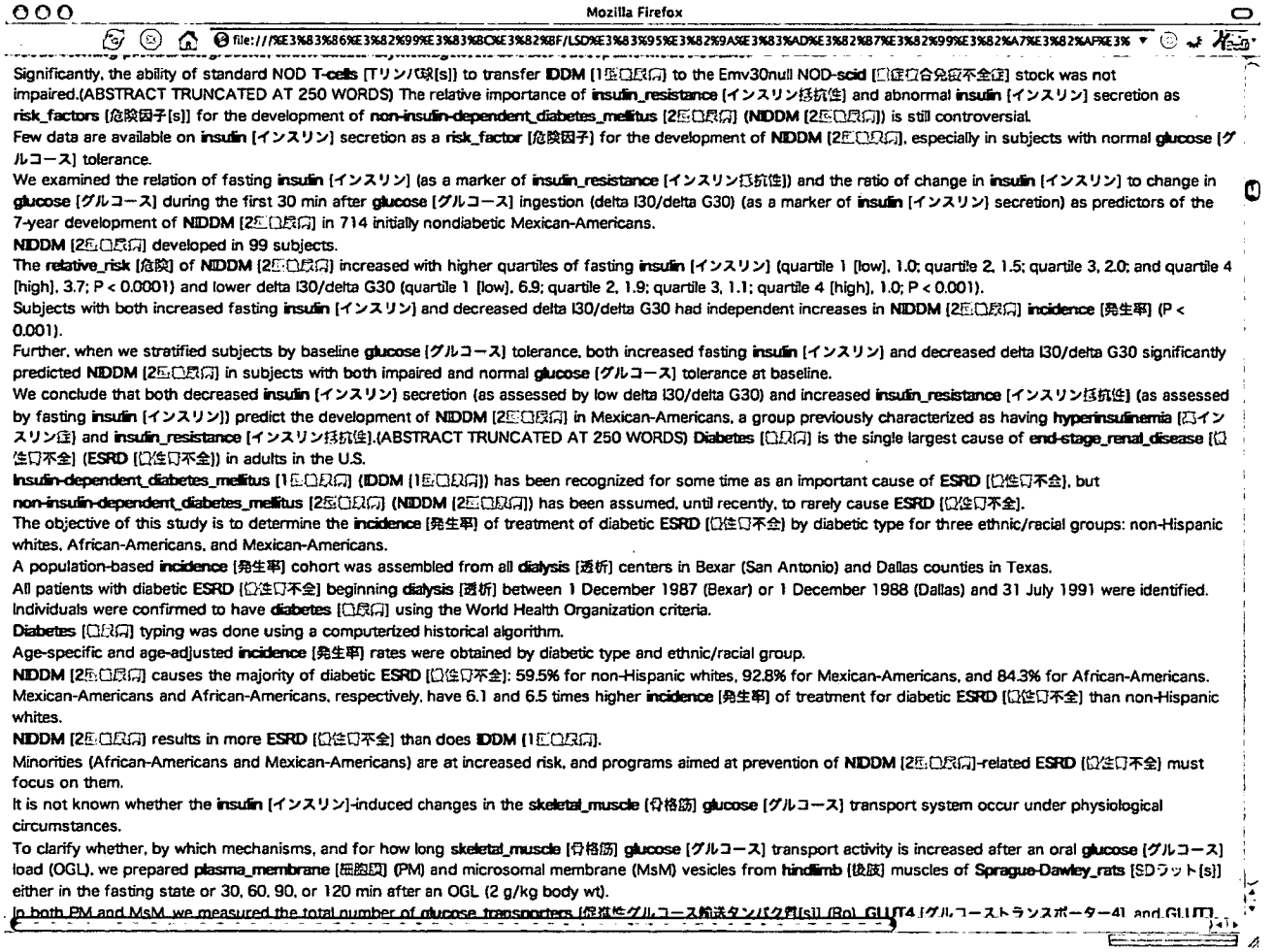


図４　PubMed抄録に出現する専門用語へのシソーラス辞書によるタグ付け

赤色は病名や症候名，青色は医薬品や生体分子名，茶色は解剖学名を表している。この例では，糖尿病に関連する用語が，略語まで含めてもれなくタグ付けされていることがわかる。現在，同様な機能を持つ日本語解析スクリプトを開発している。

糖尿病性腎症	糖尿病性網膜症	糖尿病性神経障害	2型糖尿病		
				disease	14
					260
慢性腎不全	集団検診	痛痛	インスリン	molecule	560
アンジオテンシン	血管内皮増殖因子A	診断	グルコース	molecule	317
蛋白尿	黄斑変性症	ストローク	危険因子	method	171
アンジオテンシン変換酵素阻害薬	盲目	ヘルペス後神経痛	有病率	method	154
危険因子	硬膜腫	ストレプトゾシン	発生率	method	83
レニン	白内障	グルコース	メトホルミン	molecule	75
グルコース	硝内障	神経成長因子	死亡率	method	53
アルブミン	黄斑浮腫	チオクト酸	食事	method	52
糖尿病	有病率	インスリン	診断	method	50
トランスフォーミング増殖因子β	危険因子	発生率	チアゾリジンジオン	molecule	44
クレアチニン	レーザ	動物モデル	トリグリセリド	molecule	42
死亡率	血管内皮増殖因子	危険因子	グルコキナーゼ	molecule	41
アンジオテンシン受容体	硝子体切除	血圧	罹患率	method	40
インスリン	写真	アルデヒド還元酵素	肥満度指数	method	39
ストレプトゾシン	発生率	ニコチンアミドアデニンジヌクレオチド	血糖	molecule	36
2型糖尿病	糖尿病	有病率	集団検診	method	35
終末糖化産物	腫瘍	糖尿病	炭水化物	molecule	34
血管内皮増殖因子A	糖尿病性腎症	βペプチド	動物モデル	method	34
透析	診断	死亡率	ロサルタン	molecule	33
高血糖	2型糖尿病	罹患率	アカルボース	molecule	32
アンジオテンシンII	網膜剥離	2型糖尿病	コレステロール	molecule	32
診断	インスリン	早期診断	Hepatocyte Nuclear Factor 1-alpha	molecule	27
組織診	網膜穿孔	細胞内シグナル分子	アミロイド	molecule	27
発生率	高血糖	パーキンソン病	PPARγ	molecule	25
肥大	増殖性硝子体網膜症	外傷性疾患	過体重	disease	25
細胞内シグナル分子	グルコース	切断	アルブミン	molecule	25
ロサルタン	死	リドカイン	血糖降下薬	molecule	24
体脂肪	網膜症	集団検診	インスリン受容体	molecule	24
罹患率	ソマトスタチン	ナトリウムチャネル	ホモシステイン	molecule	24
有病率	1型糖尿病	抗痙攣薬	リポタンパク質	molecule	24

図5　PubMed 抄録（1.6 GB）でのシノニム共起解析データ例

糖尿病に関連する4つの病名と1文中で共起する上位30概念ランキング

D. 考察

医療情報化社会において医療等の安全を達成するためには，市販後の医療情報や調査データを解析することによる有害作用の知識発見を早期に，確実かつ網羅的に行う必要性がある。そのような医療情報のほとんどは，文章（テキスト）として記述される。医療現場において今後，急速に電子化が推進され，大量のテキスト情報が発生すると予想できる。しかし医療情報を記述する用語については，病名，医薬品名などで国際的協調によって表記の統一化の努力が続けられているが，実際に FDA 等で公開されている医療テキストを解析すると，様々な用語が統一されずに用いられていることが過去の調査研究から分かっている。我が国において状況はさらに深刻であり，医薬品副作用報告でMedDRA/Jなどの規制用語が用いられているものの，それ以外の医療文書においては日本語では英語以上に多種多様の用語が使われているのが実態である。これまでの用語集は表記や分類を統一する方針で制作されており，網羅性に問題がある。医療情報の解析を行うためには，自然言語処理によって英語と日本語の語彙を網羅し，かつ事物（医薬品等）や概念（病名等）の同義性や関連性をツリー状に整理したオントロジーを構築することが最優先の課題と考えた。

テキストマイニング技術をゲノム科学に応用し，遺伝子と発現プロファイルや代謝パスウェイとの関係から創薬標的の発掘や遺伝子の機能推定を行おうとする研究は情報科学の領域で盛んに行われ，一部は商品化もされている。しかし，テキストマイニングを副作用情報の発見に応用しようとする試みはほとんどない。機能が不明な遺伝子の機能推定とは異なり，薬品名や症候名は（表記は統一されていないものの）限られた数の

語彙から構成されており，本来コンピュータによるテキストマイニング処理には適した材料である。しかしながら，過去に誰も着手しなかった最大の原因は，あらゆる文書での日本語・英語を網羅する「辞書」が存在しなかったためと考えられる。事実，医薬品名，遺伝子，病名の各用語はそれぞれ独自に国内外で規定されているが，これらの相互関係を多様なボキャブラリを含めて網羅的に記述したデータベースは今なお皆無である。専門領域に特化したオントロジーの構築は，検索エンジン技術を発展させる研究として情報科学でもきわめて注目されており，本研究で構築したLSDシソーラスは，医学オントロジーのプロトタイプとして有用な資源であると考えられる。

本研究によって，臨床現場から発生する大量の電子化された生の文書を早期に定量的に分析し，有害事象の早期発見を可能にするシステムの開発が可能性を帯びてきた。すなわち，様々な医療情報から有害事象が疑われるレポートを自動抽出し，人間による最終的な知識発見を支援する実用システムが完成されよう。国内での使用にあたっては，日本語解析のためのスクリプトを今後開発する必要があるが，辞書にはすでに日本語を十分量収録しているため，形態素解析などの複雑な方策を用いなくとも，関係抽出までは問題なく処理できると予想できる。今後は，シソーラスに収録した解剖部位などの場所情報や方法・技術などのデバイス情報を関係抽出に役立てる工夫が求められるだろう。

また，抽出される薬物（A）と有害事象（B）との関係の情報抽出に留まらず，DNAアレイを用いた他の研究における薬物（A）と遺伝子（C）との関係と論理的に組み合わせることによって，有害事象（B）と遺伝子（C）との関係が示唆されることになり，副作用メカニズムを実験科学で立証するための着眼点が提案されうる。これらの諸観点から，本研究の成果はIT医療時代にあって極めて高い実現性と波及効果が期待される。

なお，本研究で構築したLSDは，辞書部分および外部情報へのリンクについては無料サービスとして京都大学で公開している。今後，シソーラスツリーと共起データを実装し，情報ポータルとしてバージョンアップさせる予定である。

E. 結論

本研究によって，医薬品と疾患，症状に加えて，関連する技術や方法，解剖部位や生物名までを網羅した頑強なシソーラス辞書がほぼ完成した。また，英語テキストについて文中でのキーワード共起解析を高速かつ簡便に行うための処理プログラムを開発し，テキストマイニングからデータマイニングへの橋渡しが可能であることを示した。

今後はさらに解析結果のフィードバックからシソーラス辞書の網羅性を高めると共に，共起解析の結果を二項関係から関連性の記述へ発展させることで医療・生命科学オントロジーの構築に向けて発展させることが期待される。その際，医薬品と作用点あるいは適応症のデータベースを別途，構築することで，有害事象の判別が可能になると考えられる。また，日本語解析プログラムの開発を行い，実際の医療文書の解析と評価を繰り返すことによって，十分な実用性と有用性を有する医療情報システム設計が可能になると結論できる。

[参考文献]

1. ライフサイエンス辞書プロジェクト編著，ライフサイエンス必須英和辞典，羊土社(2005)
2. 金子周司，ライフサイエンス辞書，医学のあゆみ，210, 1062-1063 (2004)
3. 藤田信之，金子周司，ライフサイエンスのための英和変換ツール，コンピュータサイエンス，2(1)，41-45 (1995)

F. 研究発表

1．論文・著作物

1. ライフサイエンス辞書プロジェクト監修，ライフサイエンス英語表現使い分け辞典，羊土社，東京，2007.
2. ライフサイエンス辞書プロジェクト監修，ライフサイエンス論文作成のための英文法，羊土社，東京，2007.

2．学会発表

1. 伊藤悦子，金子周司．分子薬理学的知識を記述する新たな三項関係データベースの開発．第27回医療情報連合大会（神戸，2007年11月）．
2. 金子周司，大武博，藤田信之．医薬品名の同義語辞書と自動分類タガーの開発．日本薬学会第128年会（横浜，2008年3月）
3. 河本健，大武博，藤田信之，鵜川義弘，竹内浩昭，竹腰正隆，金子周司．ライフサイエンス辞書：英語での研究論文作成を支援する辞書システム−第4報−．広島大学歯学会（広島，2007年6月）

G. 知的財産権の出願・登録状況（予定も含む）

1．特許取得

本システムについては，医療機関などにおいて実用的なシステム完成を目指しており，特許取得よりむしろ無料配布を考えていきたい。

2．実用新案登録

なし

3．その他

＜他機関公開サービスへのデータ提供＞

1. 情報学研究所バイオポータル Jabion
   http://www.bioportal.jp/Gene_search/search/search.cgi
2. 理化学研究所統合データウェルス OmicScan
   http://omicspace.riken.jp/PosMed/search
3. 京都大学化学研究所バイオインフォマティクスセンターKEGG
   http://kegg.jp/
4. 情報通信研究機構 NICT 言語グリッドプロジェクト
   http://langrid.nict.go.jp/jp/
5. 統合データベースプロジェクト（公開準備中）
   http://lifesciencedb.jp/
6. 東京大学生命科学構造化センター CSLS Search（公開準備中）
   http://www.csls.c.u-tokyo.ac.jp/csls-search

厚生労働科学研究費補助金（医療安全・医療技術評価総合研究事業）
分担研究報告書

医療情報解析のためのテキストマイニングエンジンの開発

分担研究者：奥野恭史（京都大学大学院薬学研究科・統合薬学フロンティア教育センター）

[研究要旨]
本研究は、医薬品の副作用（有害事象）のレポートや医療情報の解析・評価に、テキストマイニング技術を適用し、薬物有害事象の情報解析システムを開発することにより、ＩＴ時代を迎える医療における効率良く確かな安全体制の実現を情報技術的に支援することを目的とする。
昨年に引き続き、本年度の分担研究としては、「ＸＭＬデータベースの構築とテキストマイニングエンジンの開発」、「薬物の主作用点データベース（GPCR－リガンド相互作用データベース）の開発」の２点の研究開発を行った。「ＸＭＬデータベースの構築とテキストマイニングエンジンの開発」としては、昨年開発したXML型データベースのJAPIC添付文書記載病名集データベースに、テキストマイニングエンジンの拡張として全文検索機能および他データベースへの検索機能を付与した。また、「薬物の主作用点データベース（GPCR－リガンド相互作用データベース）の開発」としては、医薬品の薬効や副作用の総合的な解析のインフォマティクス基盤として開発した薬物とタンパク質との相互作用データベースGLIDA（http://pharminfo.pharm.kyoto-u.ac.jp/services/glida）の医薬品エントリーの大幅増加と検索機能の拡張を行った。

A. 研究目的

本研究は、医薬品の副作用（有害事象）のレポートや医療情報の解析・評価に、テキストマイニング技術を適用し、薬物有害事象の情報解析システムを開発することにより、ＩＴ時代を迎える医療における効率良く確かな安全体制の実現を情報技術的に支援することを目的とする。

B. 研究方法

１．ＸＭＬデータベースの構築とテキストマイニングエンジンの開発

薬物有害事象の自動抽出を目的としたテキストマイニングエンジンの開発素材として、（財）日本医薬情報センター（JAPIC）の添付文書記載病名集を用いた。JAPIC添付文書記載病名集は、医薬品の薬効や副作用情報など本研究対象に必要な情報が記載されており、さらにXML形式での電子データが供給されている。そこで、昨年、XMLデータベースNeoCoreを用いて、JAPIC添付文書記載病名集からの用語の自動抽出、構造化を行い、JAPIC添付文書記載病名集を包含するテキストマイニング用データベースの構築を行った。本年は、テキストマイニングエンジンの開発として、全文検索アルゴリズムを実装するとともに、PubMedおよびGoogleへの同時検索機能の開発も行った。

2．薬物の主作用点データベース（GPCR－リガンド相互作用データベース）の開発

医薬品による副作用の総合的な解析には、薬物の作用点となる標的タンパク質との相互作用様式を情報学的に処理する基盤技術の整備が必須となる。本研究では、薬物の主作用点データベースとして GPCR－リガンド相互作用データベース（GLIDA データベース）を構築し、http://pharminfo.pharm.kyoto-u.ac.jp/services/glida より公開している。市販医薬品の大半は、G タンパク質共役型受容体(GPCR)を薬物作用点にしていることから、本データベースがプロトタイプになり得る。GLIDA は、GPCR のバイオ情報、リガンドのケミカル情報、および GPCR とリガンドの相互作用情報の3種類の情報より構成される。GPCR のエントリはヒト、ラット、マウスに限定し、バイオ情報は GPCRDB から取得した。また、GPCR と結合するリガンドのエントリとそのケミカルデータ（化学名、構造式、分子量, MDL Mol ファイルなど）は IUPHAR Receptor Database, PubMed, PubChem および MDL ISIS などの公共または商用のデータベースから取得した。本年は、医薬品情報として、GPCR のリガンドエントリーの大幅増加を図るとともに、そのリガンド検索機能の改良を行った。

なお、本研究は計算機によるシステム開発であり、倫理面に関する問題は一切無い。

C. 研究結果

1．XMLデータベースの構築とテキストマイニングエンジンの開発

昨年度に、テキストマイニング用データベースとして開発した JAPIC 医薬品添付文書記載病名集データベースを対象に、全文検索機能の開発を行った。

全文検索エンジンとしては、XML データベース NeoCore のパッケージである Quick Solution を用いてその実装を行った。これにより、JAPIC 医薬品添付文書記載病名集の高速全文検索が可能となり、現在、副作用情報等の医療情報の収集を行っている。

また、更なる医療情報の収集を目的として、本データベースと他の公共データベースとの連携機能の開発を行った。具体的には、代表研究者が開発してきたライフサイエンス辞書との連携を図ることにより、日本語と英語の自動相互変換を行い、PubMed および Google への同時検索を可能にした。

2．薬物の主作用点データベース（GPCR－リガンド相互作用データベース）の開発

本研究では、薬物と GPCR の相互作用データベースとして開発、公開している GLIDA の医薬品情報の大幅増加を行った。具体的には、24,077 件ものリガンドエントリー、および 39,140 件ものリガンドーGPCR の相互作用エントリーの登録に至っている。これは、昨年度公開時点の約 35 倍と 20 倍ものエントリーの増加にあたり、公共の医薬品データベースとしては世界最大級のエントリーを誇っている。

また、リガンドエントリーの大幅増加に伴い、リガンド検索ツールの大幅改良も行った。以前のリガンド分類は階層型クラスタリングに基づくものであったが、計算量の問題から、数万エントリーにはこのクラスタリングは適さない。そこで、本研究では、主成分分析（PCA）に基づいて、全リガンドエントリーの分類を行った。さらに、リガンドの部分構造に基づく類似化合物検索エンジンの開発と実装を行い、一般ユーザーからのデータベース検索を可能にしている。

なお、本データベース GLIDA は http://pharminfo.pharm.kyoto-u.ac.jp/services/glida より公開している。

D. 考察

1．XMLデータベースの構築とテキストマイニングエンジンの開発

本研究において開発している基幹システムであるXMLデータベースは、非常に汎用性に富んだシステムである。従って、現在登録中のJAPIC医薬品添付文書記載病名集はデータの一例であり、今後他の文書データを登録し、総合的なテキストマイニングを展開する。

また、本研究で実装した全文検索エンジンを活用して、医薬品添付文書内に収録されている副作用情報の検索と分析を展開する。

2．薬物の主作用点データベース（GPCR－リガンド相互作用データベース）の開発

GLIDAデータベースは市販の医薬品の半分以上の標的分子となっているGPCRとそれに作用する薬物の相互作用に関する知識データベースであるとともに、その相互作用メカニズムの解明に関する知識を提供し得るケミカルゲノミクスのためのデータベースである。医薬品の薬効や副作用の総合的な解析には、薬物の作用基点となる遺伝子との相互作用様式を情報学的に処理する基盤技術の整備は必須であり、本データベース構築によりその基盤は確立された。今後、GLIDAデータベースに登録された医薬品情報と、上述、テキストマイニングデータベースとの連携を図る。

E. 結論

1．XMLデータベースの構築とテキストマイニングエンジンの開発

JAPIC添付文書記載病名集データベースのテキストマイニングエンジンの開発として、全文検索アルゴリズムを実装するとともに、PubMedおよびGoogleへの同時検索機能の開発も行った。

2．薬物の主作用点データベース（GPCR－リガンド相互作用データベース）の開発

医薬品の薬効や副作用の総合的な解析のインフォマティクス基盤として、薬物とタンパク質との相互作用データベースGLIDAの大幅エントリー増加と検索機能の拡張を行った。これらは、http://pharminfo.pharm.kyoto-u.ac.jp/services/glidaより公開している。

F. 研究発表

1．論文発表

1. Niijima, S. and Okuno, Y. "Laplacian Linear Discriminant Analysis Approach to Unsupervised Feature Selection" *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, in press

2. Okuno, Y., Tamon, A., Yabuuchi, H., Niijima, S., Minowa, Y., Tonomura, K., Kunimto, R. and Feng, C. "GLIDA: GPCR-Ligand Database for Chemical Genomics Drug Discovery – Database and Tools Update" *Nucleic Acids Research*, **36**, D907-12, 2008

3. Kitajima, M., Minowa, Y., Matsuda, H. and Okuno, Y. "Compound-Transporter Interaction Studies using Canonical Correlation Analysis" *Chem-Bio Informatics J.*, **7**, 24-34, 2007

4. Yamamoto, H., Takematsu, H., Fujinawa, R., Naito, Y., Okuno, Y., Tsujimoto, G., Suzuki, A. and Kozutsumi, Y. "Correlation index-based responsible-enzyme gene screening (CIRES), a novel DNA microarray-based method for glycan biosynthesis enzyme gene" *PLoS ONE*, **2**, e1232, 2007

5. Ikeda, A., Miyazaki, T., Kakizawa, S., Okuno, Y., Tsuchiya, S., Myomoto, A., Saito, S. Y., Yamamoto, T., Yamazaki, T., Iino, M., Tsujimoto, G., Watanabe, M. and

Takeshima, H. "Abnormal features in mutant cerebellar Purkinje cells lacking junctophilins." *Biochem. Biophys. Res. Commun.*, **363**, 835-9, 2007

6. Yamazaki, T., Sasaki, N., Nishi, M., Yamazaki, D., Ikeda, A., Okuno, Y., Komazaki, S., and Takeshima, H. "Augmentation of drug-induced cell death by ER protein BRI3BP." *Biochem. Biophys. Res. Commun.*, **362**, 971-5, 2007

2．学会発表

1. 日本薬学会128年会　日本薬学会奨励賞受賞講演「バイオ空間とケミカル空間の包括的相関解析とそのインシリコ創薬への研究展開」2008年3月28日

2. 第2回ClassA システムバイロジーセミナーin 東京「ケミカル・スペースでの化合物探索」2008年2月28日

3. 第2回ClassA システムバイロジーセミナーin 関西「ケミカル・スペースでの化合物探索」2008年2月25日

4. 第3回三重ゲノム創薬フォーラム「薬学研究におけるアレイインフォマティクス」2008年2月15日

5. 平成19年度 第2回産業情報交流会「ケミカルゲノミクスに基づくインシリコ化合物探索他」2007年10月22日

6. 新産業を創る先端科学技術フォーラム 2007「ポストゲノム創薬のための新技術」セッション「ケミカルゲノミクスに基づく創薬インフォマティクス」2007年10月18日

7. 第66回日本癌学会学術総会 International Session「Chemical Genomics for Cancer Research」「Knowledge Discovery and Data Mining in Chemical Genomics」2007年10月3日

8. バイオビジネスステーション　卒業生交流会2007年7月21日

G. 知的財産権の出願・登録状況（予定も含む）

1．特許出願

1. 特願2007-53322、「マイクロRNA標的遺伝子予測装置」、平成19年3月2日出願、出願人　東レ株式会社、発明者　奥野恭史、辻本豪三、国本亮、寺澤和哉、土屋創健、秋山英雄、妙本明

2. 公開番号WO2007/004479A; 特開2007-11752、「データ処理装置、データ処理プログラム、それを格納したコンピュータ読み取り可能な記録媒体、およびデータ処理方法」、平成19年1月11日公開、出願人　京都大学、発明者　奥野恭史、辻本豪三、梁智允、種石慶

3. 公開番号WO2007/139037A1; PCT/JP2007/060736 特願2006-147433、「ケミカルゲノム情報に基づく、タンパク質－化合物相互作用の予測と化合物ライブラリーの合理的設計」、平成18年5月26日（国内）平成19年5月25日（国際）出願、出願人　京都大学、発明者　奥野恭史、種石慶、辻本豪三

2．実用新案登録

無し

3．その他

無し

研究成果の刊行に関する一覧表

書籍

著者氏名	書籍全体の編集者名	書籍名	出版社名	出版地	出版年
金子周司ほか	ライフサイエンス辞書プロジェクト監修	ライフサイエンス英語表現使い分け辞典	羊土社	日本	2007
金子周司ほか	ライフサイエンス辞書プロジェクト監修	ライフサイエンス論文作成のための英文法	羊土社	日本	2007
金子周司ほか	今堀和友・山川民夫監修	生化学辞典第4版	東京化学同人	日本	2007
奥野恭史ほか	石渡信一・桂勲・桐野豊・美宅成樹 編	生物物理学ハンドブック	㈱朝倉書店	日本	2007

雑誌

発表者氏名	論文タイトル名	発表誌名	巻号	ページ	出版年
Nakao, K., Shirakawa, H., Sugishita, A., Matsutani, I., Niidome, T., Nakagawa, T., and Kaneko, S.	Ca^{2+} mobilization mediated by transient receptor potential canonical 3 is associated with thrombin-induced morphological changes of 1321N1 human astrocytoma cells.	**J. Neurosci. Res.**	in press		2008
Niijima, S. and Okuno, Y.	Laplacian Linear Discriminant Analysis Approach to Unsupervised Feature Selection	**IEEE/ACM Transactions on Computational Biology and Bioinformatics**	in press		2008
Okuno, Y., Tamon, A., Yabuuchi, H., Niijima, S., Minowa, Y., Tonomura, K., Kunimto, R. and Feng, C.	GLIDA: GPCR-Ligand Database for Chemical Genomics Drug Discovery – Database and Tools Update	**Nucleic Acids Research**	36	D907-12	2008

Nakagawa, T., and Kaneko, S.	Neuropsychotoxicity of Abused Drugs: Molecular and Neural Mechanisms of Neuropsychotoxicity Induced by Methamphetamine, 3,4-Methylenedioxymethamphetamine (Ecstasy), and 5-Methoxy-N,N-diisopropyltryptamine (Foxy).	**J. Pharmacol. Sci.**	106	2-8	2007
Yamauchi, Y., Izumi, T., Unemura, K., Uenishi, Y., Nakagawa, T., and Kaneko, S.	Acceleration of serotonin transporter transport-associated current by 3,4-methylenedioxymethamphetamine (MDMA) under acidic conditions.	**Neurosci. Lett.**	428	72-76	2007
Tanaka, K., Shirakawa, H., Okada, K., Konno, M., Nakagawa, T., Serikawa, T., Kaneko, S.	Increased Ca^{2+} channel currents in cerebellar Purkinje cells of the ataxic groggy rat.	**Neurosci. Lett.**	426	75-80	2007
Tokuda, S., Kuramoto, T., Tanaka, K., Kaneko, S., Takeuchi, I.K., Sasa, M., Serikawa, T.	The ataxic groggy rat has a missense mutation in the P/Q-type voltage-gated Ca^{2+} channel $alpha_{1A}$ subunit gene and exhibits absence seizures.	**Brain Res.**	1133	168-177	2007
Kitajima,M., Minowa, Y., Matsuda, H. and Okuno, Y.	Compound-Transporter Interaction Studies using Canonical Correlation Analysis	**Chem-Bio Informatics J.**	7	24-34	2007
Yamamoto, H., Takematsu, H., Fujinawa, R., Naito, Y., Okuno, Y., Tsujimoto, G., Suzuki, A. and Kozutsumi, Y.	Correlation index-based responsible-enzyme gene screening (CIRES), a novel DNA microarray-based method for glycan biosynthesis enzyme gene	**PLoS ONE**	2	e1232	2007

Ikeda, A., Miyazaki, T., Kakizawa, S., Okuno, Y., Tsuchiya, S., Myomoto, A., Saito, SY., Yamamoto, T., Yamazaki, T., Iino, M., Tsujimoto, G., Watanabe, M. and Takeshima, H.	Abnormal features in mutant cerebellar Purkinje cells lacking junctophilins.	**Biochem. Biophys. Res. Commun.**	363	835-9	2007
Yamazaki,T., Sasaki, N., Nishi, M., Yamazaki, D.,Ikeda,A.,Okuno,Y.,Komazaki,S.,andTakeshima, H.	Augmentation of drug-induced cell death by ER protein BRI3BP.	**Biochem. Biophys. Res. Commun.**	362	971-5	2007

分子薬理学的知識を記述する新たな三項関係データベースの開発

伊藤 悦子　金子 周司
京都大学大学院薬学研究科

The development of novel drug database describing molecular pharmacological knowledge

Ito Etsuko Kaneko Shuji
Graduate School of Pharmaceutical Sciences, Kyoto University

We have developed a novel drug database describing sites and modes of action from molecular pharmacoloical point of view. About 18,000 drug names were collected from JAPIC, JAN, ATC, KEGG and MeSH databases and rearranged to a synonym table listing 2,800 active chemical compounds. The sites of action were assigned to one or few target molecules described with MeSH descriptors and RefSeq genes. The mode of action for each interaction was described with simle descriptors, such as 'inhibition' and 'activation'. As a result, the modes of drug ation were assigned to approximately 70% of active compounds to 500 target molecules. The molecular pharmacology database will be of useful for constructing medical ontology.

Keywords: pharmacological action, ligandp, ontology, drug name

1. はじめに

生命情報学の発展により,生体分子や化合物情報の網羅的なデータベースが整いつつある。これら物質データベースが完成の域へ近づくにつれ,それぞれの関係や相互作用を記述する統合的なデータベースが知識基盤として構築されつつある[1,2]。

医薬品の作用は,究極的には活性成分である化学物質と生体分子との相互作用に単純化できる[3]。医薬品の全身作用を収録する添付文書や副作用情報は電子化が進んでいるが,薬物の分子作用点や作用メカニズムに関する薬理学的知識について,単純化した相互作用として整理したデータベースは数多くない[4,5]。特に,日本語で記述される製剤名や商品名などの多様な表記をカバーし,医療文書のテキストマイニングに耐えうる網羅的な記述子を備えた薬理作用データベースはまだない。

本研究ではまず,医薬品として用いられる化学物質に対して付与される製剤名や商品名に加え,生体分子を記述する際に用いられる略語や省略形まで含め,それらの英語と日本語表記を網羅するシノニムテーブルを構築した。次にこれらを用いて,化学物質と生体分子の相互作用を数少ない相互作用様式とともに三項関係によって記述した新たな分子薬理学データベースを構築した。

2. 方法

JAPIC医療用医薬品集および日本医薬品一般名称(JAN)より抽出した商品名,製剤名,有効成分名を基本にして,欧米で用いられる名称および表記をATC,KEGG,MeSHより追加した。こうして集めた日本語および英語の名称について,有効成分が同一なものの集合を同義語として整理した。漢方製剤は有効成分が明らかな場合以外は除外した。MeSHに収録されている場合は,Descriptor表記を代表記述子とした。次に,ライフサイエンス辞書(LSD)[6]に収録されている英和対訳の物質名をMeSHツリーと関連づけ,生体分子名の同義語テーブルを構築した。

作用点の定義には,国内外の薬理学テキストと独自に収集したPubMedコーパスの解析結果を参考にした。これらテキスト(500 MB)に出現する医薬品および生体分子にタグを付与し,続いてタグ同士の共起頻度をperlスクリプトで解析することによって相互作用と考えられるペアを抽出した。この化合物と標的生体分子との相互作用について,様式を表現する阻害,活性化などの少数の記述子を用いて三項関係として記述した。標的分子が明らかでない場合は,細胞ないし組織レベルにおいて報告されている知識を,別の関係テーブルにおいて同様にMeSH用語と少数の作用様式を用いて整理した。以上のデータはFileMaker Proを用いたリレーショナルデータベースとした。

3. 結果

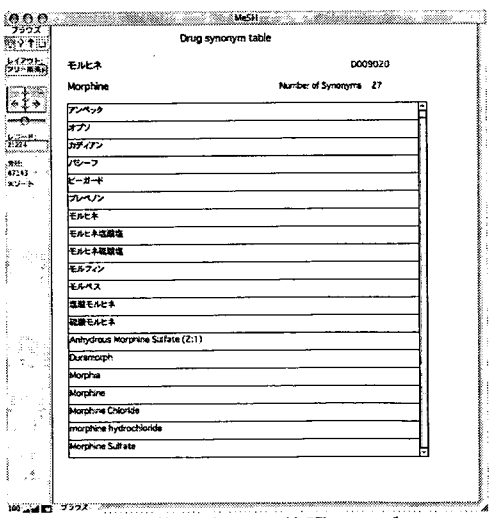


図1 医薬品名の同義語テーブル

収集した医薬品名称(商品名,製剤名,有効成分名)は約18,000種類(うち,日本語は約6,000)であった。これらを有効成分によって整理し,2,842種類の有効成分リストを作成した(図1)。このうちMeSHには2,545種類,ATCには2,599種類が収録されており,どちらにも収録されていない成分はなかった。次に,各有効成分について,一般的に行われている薬効分類を適用し,化合物群によるグループ化を行った。一方,LSDから抽出した生体分子名は英和対訳として19,595種類であった。これらをMeSHツリーを参考にして同義語テーブルとして整理した結果,4,272種類の標的候補分子リストを作成した。

上述の医薬品名および標的候補分子名が薬理学教科書およびPubMedコーパスのテキスト中で共起する頻度をカウントした結果,ほとんどの医薬品について共起頻度の上位ペアに主作用点が浮かび上がった(図2)。そこで,このデータを参考にして,個々の薬物について薬理作用点である標的分子を特定した。この際,同一の薬効分類に属する医薬品について,同一の作用点を有するかのチェックを行った。

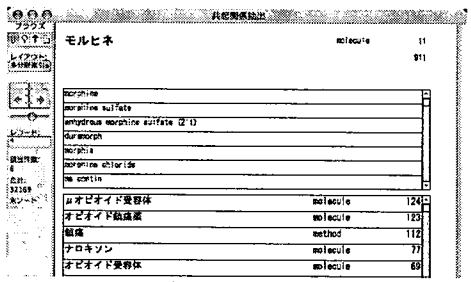


図2 化合物と生体分子の関係抽出

各々の相互作用様式については,標的分子が受容体,酵素,膜輸送タンパク質といったタンパク質の場合には,その結合様式が非共有結合であっても共有結合であっても,結果としてそのタンパク質の機能に与える影響を阻害,活性化など少数の記述子を用いて表現した(図3)。

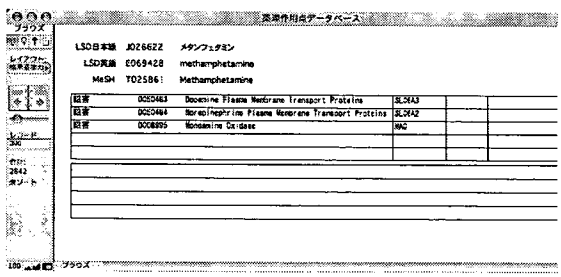


図3 分子作用点テーブル

また,転写調節に影響する薬物の場合は,主たる標的である結合タンパク質とともに,結果的に発現が調節される遺伝子のうち,主たる薬効の発揮に関与していると考えられる標的タンパク質についても発現上昇,発現抑制などの記述子とともに収録した。同様に,ホルモン補充療法などの場合,類似した薬理作用をもつ生体成分の補充であることを示すとともに,生体成分が作用する標的を記述した。これらの収録にあたっては,医薬品名にMeSHおよびLSDでのコード番号とともに,生体分子にはMeSH Descriptor IDおよびRefSeq遺伝子名を付与し,他データベースへの互換性と拡張性を持たせた。

以上の結果,全体の70%に相当する薬物について,計500種類の標的分子との特異的な相互作用を収録した。分子レベルの作用が不明瞭で,細胞ないし臓器レベルでの効果が知られている残りの薬物については,それらを別テーブルにおいてMeSH用語と相互作用記述子を用いて収録した。

4. 考察

約2,800種類の医薬品有効成分のうち,70%に分子作用点が記述でき,その作用点が500種類であったという結果は,過去に調べられた医薬品の標的についての調査とよく一致している[7,8]。標的分子が1つ以上に及ぶ薬物は約1,000種類と多く,薬効を考える上で必要な知識を網羅していると考えられる。しかしながら,MeSH収録語を中心とした記述では一部において表現できないサブタイプが存在することが明らかになった。また,今回は相互作用を定性的にのみ表現したが,実際のテキスト解析に応用する場合を想定すると,複数の作用点に異なる親和性をもって作用する場合などで定量的な尺度を導入する必要性が感じられた。

医薬品の同義語関係を整理するとともに分子作用点を正規化した本データベースは,将来的に医薬品の適応症や副作用,あるいは代謝パスウェイ等のデータベースと組み合わせることによって有害作用の予測や因果関係の解析に有用な資源となると考えられる。

参考文献

[1] KEGG生命システム情報統合データベース.http://www.kegg.jp/.京都大学化学研究所.
[2] OmicSpace.http://omicspace.riken.jp.理化学研究所.
[3] 金子周司(辻本豪三,田中利男編).創薬統合データベース.21世紀の創薬科学.共立出版, 1998: 141.
[4] Wishart DS et al.DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nuc. Acids Res. 2006; 34: D668.
[5] 吉川澄美,松村和美,小長谷明彦.薬機能オントロジー:分子機能と生体現象の関連化.人工知能学会第5回セマンティックウェッブとオントロジー研究会抄録集, 2004.3.
[6] ライフサイエンス辞書.http://lsd.pharm.kyoto-u.ac.jp/.京都大学薬学研究科.: 1960.
[7] Drews J.Drug discvery: a historical perspective. Science 2000; 287: 1960.
[8] Hopkins AL and Groom CR.The druggable genome. Nature Rev. Drug Discov. 2002; 1: 727.

Published online 5 November 2007

Nucleic Acids Research, 2008, Vol. 36, Database issue ***D907–D912***
doi:10.1093/nar/gkm948

GLIDA: GPCR—ligand database for chemical genomics drug discovery—database and tools update

Yasushi Okuno[1,*], Akiko Tamon[2], Hiroaki Yabuuchi[1], Satoshi Niijima[1], Yohsuke Minowa[1], Koichiro Tonomura[1], Ryo Kunimoto[1] and Chunlai Feng[1]

[1]Department of PharmacoInformatics, Center for Integrative Education of Pharmacy Frontier, Graduate School of Pharmaceutical Sciences, Kyoto University and [2]Bio Science Group, IT Solution Div.1, Industry Solution Business Unit, Mitsui Knowledge Industry, Osaka city, Japan

Received September 6, 2007; Revised October 14, 2007; Accepted October 15, 2007

ABSTRACT

G-protein coupled receptors (GPCRs) represent one of the most important families of drug targets in pharmaceutical development. GLIDA is a public GPCR-related Chemical Genomics database that is primarily focused on the integration of information between GPCRs and their ligands. It provides interaction data between GPCRs and their ligands, along with chemical information on the ligands, as well as biological information regarding GPCRs. These data are connected with each other in a relational database, allowing users in the field of Chemical Genomics research to easily retrieve such information from either biological or chemical starting points. GLIDA includes a variety of similarity search functions for the GPCRs and for their ligands. Thus, GLIDA can provide correlation maps linking the searched homologous GPCRs (or ligands) with their ligands (or GPCRs). By analyzing the correlation patterns between GPCRs and ligands, we can gain more detailed knowledge about their conserved molecular recognition patterns and improve drug design efforts by focusing on inferred candidates for GPCR-specific drugs. This article provides a summary of the GLIDA database and user facilities, and describes recent improvements to database design, data contents, ligand classification programs, similarity search options and graphical interfaces. GLIDA is publicly available at http://pharminfo.pharm.kyoto-u.ac.jp/services/glida/. We hope that it will prove very useful for Chemical Genomics research and GPCR-related drug discovery.

INTRODUCTION

The family of G-protein coupled receptors (GPCRs) represents one of the most important classes of pharmaceutical targets (1). Among the more than 1000 GPCRs encoded in the human genome, more than 400 are of potential therapeutic interest (2). Currently the drugs available on the market address only 30 GPCRs, which represent a small fraction of the GPCR target family. A large majority of human-derived GPCRs still remain promising drug targets, and thus a key goal of GPCR research related to drug design is to identify new ligands for such target GPCRs.

With the unprecedented accumulation of genomic information, databases and bioinformatics have become essential tools to guide GPCR research (3). The GPCRDB (2) and IUPHAR receptor database (IUPHAR-RD) (4) are representatives of widely used public databases covering GPCRs. These databases, which provide substantial data on the GPCR proteins and pharmacological information on receptor proteins containing GPCRs, are mainly focused on biological aspects of the GPCR gene products or proteins. In spite of the significance of ligand compounds as drug leads, the relationships between GPCRs and their ligands and/or chemical information on the ligands themselves are not yet fully covered.

On the other hand, there is increasing interest in publicly collecting and applying chemical as well as biological information in the post-genome era (5–8). This new trend is called 'Chemical Genomics', and it aims to identify all possible chemical ligands and drugs for all targets families (9,10). There is a vast amount of information on the interactions between small molecules and proteins/genes. However, compound–protein interactions have not yet been analyzed on a large scale, and there are no effective methods to extract meaningful

*To whom correspondence should be addressed. Tel: +81 75 753 4559; Fax: +81 75 753 4544; Email: okuno@pharm.kyoto-u.ac.jp

information from the data in a comprehensive manner. Therefore, we need to integrate chemoinformatics and bioinformatics into a common computational platform for mining of Chemical Genomics data (11).

GLIDA (GPCR-Ligand DAtabase) is a public GPCR-related Chemical Genomics database designed to simultaneously mine biological information on GPCRs and chemical information on their ligands. It provides various analytical data regarding GPCR–ligand correlations by incorporating bioinformatics and chemoinformatics techniques, and thus it should prove very useful for GPCR-related drug discovery from the viewpoint of Chemical Genomics research. There have been several major improvements to GLIDA since it was last described in Ref. (12): (i) there are more increments in the entries of the ligands and the corresponding ligand–GPCR pairs; (ii) the ligands are originally classified using a new strategy; (iii) additional options are available within the similarity search program for the GPCRs and ligands and (iv) the graphical interface to display the correlation maps between GPCRs and ligands has been enhanced.

DATA CONTENTS

GLIDA contains three types of primary data: biological information on GPCRs, chemical information on their ligands and information on binding of the GPCR–ligand pairs. The GPCR entries were acquired from human, mouse and rat entries deposited in the GPCRDB because these three species include sufficient information regarding ligands, and rats and mice are representative model animals used in drug discovery research. The ligand-binding information was manually collected and curated using various public web sites and commercial databases such as the IUPHAR-RD, PubMed (5), PubChem (5), DrugBank (13), Ki Database (14) and MDL ISIS/Base 2.5. Table 1 indicates the size and scope of the GLIDA database. In particular, we have dramatically expanded the entry number of ligands and the corresponding ligand–GPCR pairs. The latest GLIDA version includes 24 077 ligand entries and 39 140 GPCR–ligand pair entries, representing nearly 35-fold and 20-fold increases, respectively, since the last publication of GLIDA in 2006. The total number of GPCR entries remains unchanged, but entries with associated ligand information have increased slightly, suggesting that it is difficult to de-orphan the GPCRs whose ligands have not yet been identified (15).

GPCR and ligand data

The database lists general information on GPCR and ligand data, respectively. The general information table listing GPCRs contains gene names, family names, protein sequences (in fasta format) and links to other biological databases, such as GPCRDB, UniProt (16), IUPHAR-RD, Entrez Gene (17) and KEGG (18). The ligand result page provides a general information table containing names, molecular structures, CAS registry numbers, formulas, molecular weights, structure files and links to PubChem, KEGG, ChEBI (8) and DrugBank that are in publicly available chemical databases.

Table 1. The current numbers of GLIDA ligands and GPCRs and their respective links.

Information item	Number of entries
GPCR entries	3738
Links to Entrez Gene	3073
Links to GPCRDB	3738
Links to UniProt	3738
Links to IUPHAR	446
Links to KEGG	595
Ligand entries	24 077
Cas registry number	2425
Molecular structure	23 216[a]
Links to PubChem	1821
Links to ChEBI	103
Links to KEGG	664
Links to DrugBank	479
Cluster number	300[b]
GPCR–ligand pair entries	39 140
GPCR entries	410
Ligand entries	24 077
Activity	
Agonist	8305
Full Agonist	2325
Partial Agonist	262
Antagonist	28 132
Inverse Agonist	116

[a]Molecular structures consist of MDL MOL files and original files converted into KEGG atom types. The numbers of MDL MOL files and KEGG-type files are 23 216 and 23 214, respectively. PCA calculation was performed for 23 214 KEGG-type files.
[b]This cluster number (300) is different from the number of the selected principal components (314). No compounds were assigned to 14 principal components.

Information on binding of GPCR–ligand pairs

The interaction information relating GPCRs to particular ligands, a key issue for GPCR-related drug discovery, is deposited in a relational database. GLIDA allows users to retrieve GPCR–ligand-binding information dynamically and continuously. When users retrieve a GPCR (or ligand) entry, its result page displays all entries showing the corresponding ligands (or GPCR) entries with their binding activity types, as well as references. The references are hyperlinked with the corresponding PubMed literature. The activity types include agonist, antagonist and full, partial or inverse agonist (Table 1). Here the detail annotations such as full, partial or inverse agonist are not finished yet. The ligands classified as agonists are possible full agonists or partial agonists. Inverse agonists can be also contained among the antagonists.

WEB INTERFACE AND APPLICATION

GLIDA is available at http://pharminfo.pharm.kyoto-u.ac.jp/services/glida/. The web interface of GLIDA includes a GPCR search page (Figure 1a) and a ligand search page (Figure 1b). Each page consists of a classification menu and a keyword search box. The users can search a GPCR (or ligand) manually using the classification tool,