KNApSAcKを用いた植物の効能メカニズム解明のための基盤構築

京都大学　化学研究所
西村陽介

H25.11.29
統合データ解析トライアル・中間激励会

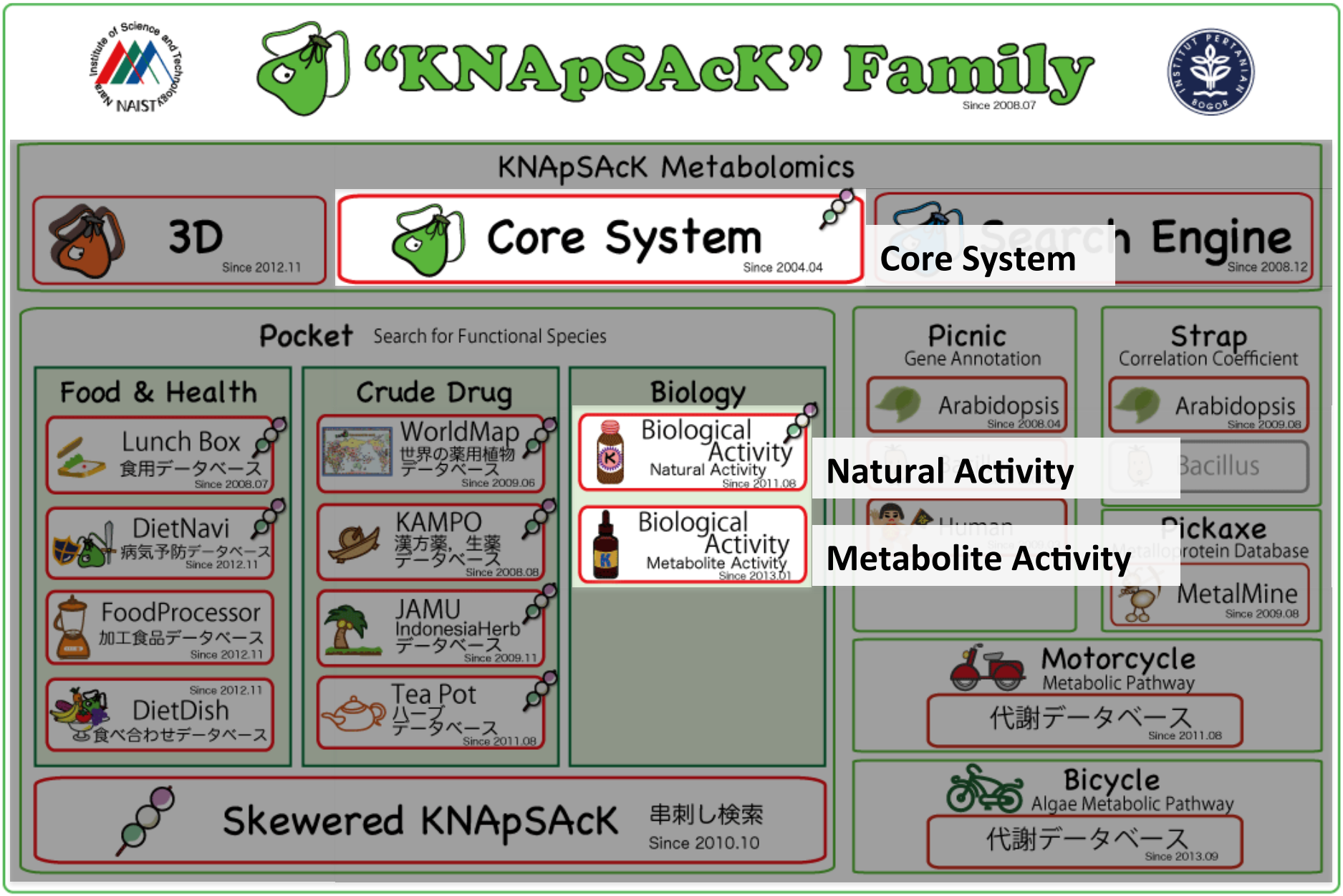
目次

- 研究開発計画の概要
- KNApSAcK
  - NCBI taxonomy を用いた生物種名の標準化と階層分類
  - 代謝産物の構造分類手法
- ChEMBL、CTDとの情報統合
  - Standard InChIやCAS RNを用いた化合物リンク作成
    - 作成されたリンクの評価
- 今後の予定
  - 相互作用タンパク質のIDマッピングと分類
  - 植物種を入力して関連する疾患リストを出力
    - webインターフェイス、APIの設計

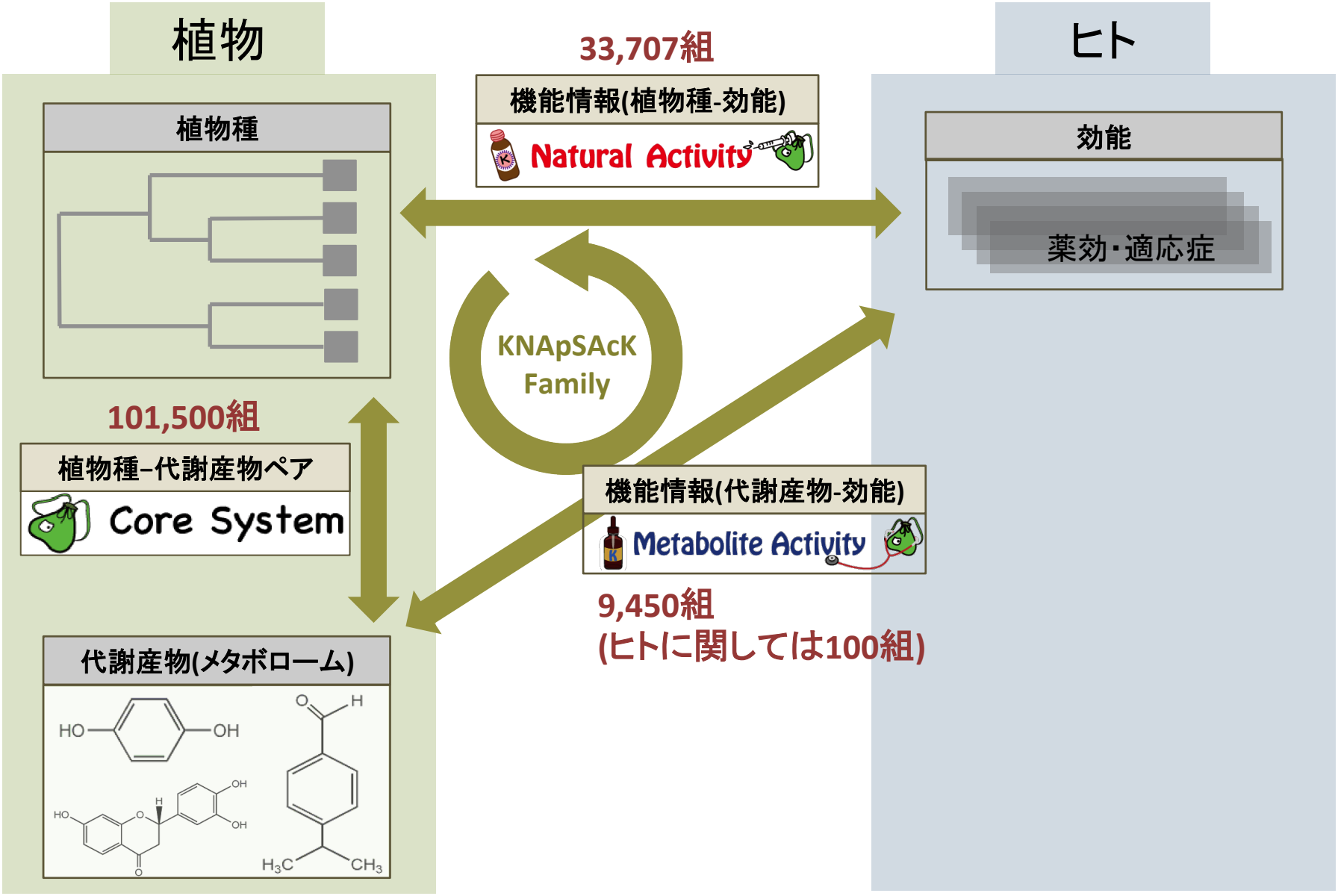
KNApSAcKと分子レベルの相互作用情報の統合による植物のヒトに対する効能メカニズムの解明

- KNApSAcK: 生物種と代謝産物の関係を体系化することを目的とした二次代謝産物データベース。
- 近年の大規模解析技術の発達に伴って、ヒトタンパク質と代謝産物の相互作用情報がデータベース上に蓄積されてきている。
- 分子レベルの相互作用情報を用いて、植物のヒトに対する効能メカニズムをより正確に解明することが植物のより良い利用につながる。

KNApSAcK Family



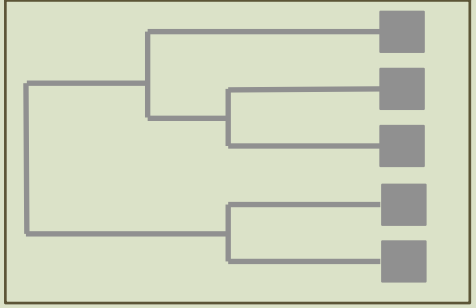
KNApSAcKの情報



KNApSAcKを利用した個体レベルの機能情報の予測

植物

植物種

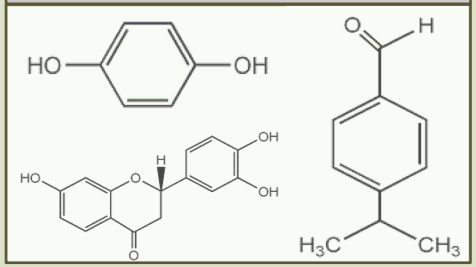


101,500組

各植物種の
メタボローム情報

Core System

代謝産物(メタボローム)



個体レベルの
機能情報

?

KNApSAcK
Family

ヒト

効能

薬効・適応症

分子レベルの
機能情報

Metabolite Activity

9,450組
(ヒトに関しては100組)

⇒他から情報を補いたい

植物代謝産物の効能を予測するためのスキーム

植物

ヒト

効能

薬効・適応症・毒性

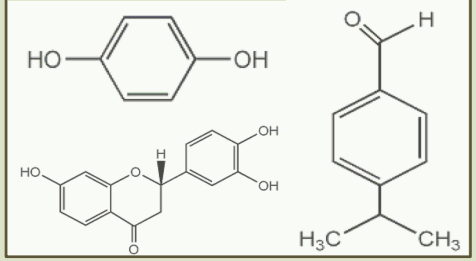
?

タンパク質機能情報

GO the Gene Ontology

KEGG PATHWAY

代謝産物(メタボローム)

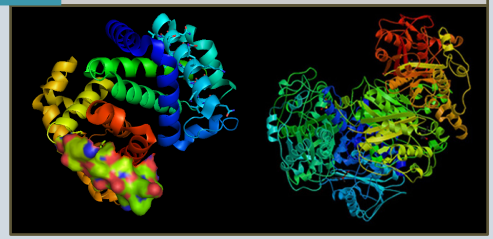


機能情報(代謝産物-タンパク)

ChEMBL

ctd

相互作用タンパク質



植物代謝産物の効能を予測するためのスキーム

植物

ヒト

スコアリング

効能

代謝産物

予測

分子レベルの機能情報

効能

薬効・適応症・毒性

タンパク質機能情報

GO the Gene Ontology

KEGG PATHWAY

代謝産物(メタボローム)

相互作用タンパク質

機能情報(代謝産物-タンパク)

ChEMBL

ctd

他のデータベースからの機能情報の追加

植物

ヒト

効能

薬効・適応症・毒性

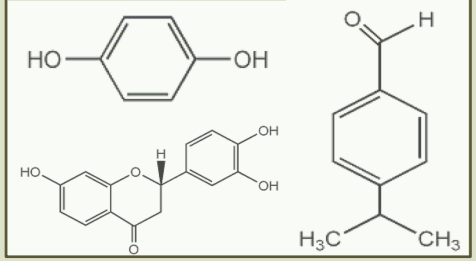
機能情報(代謝産物-効能)

ctd

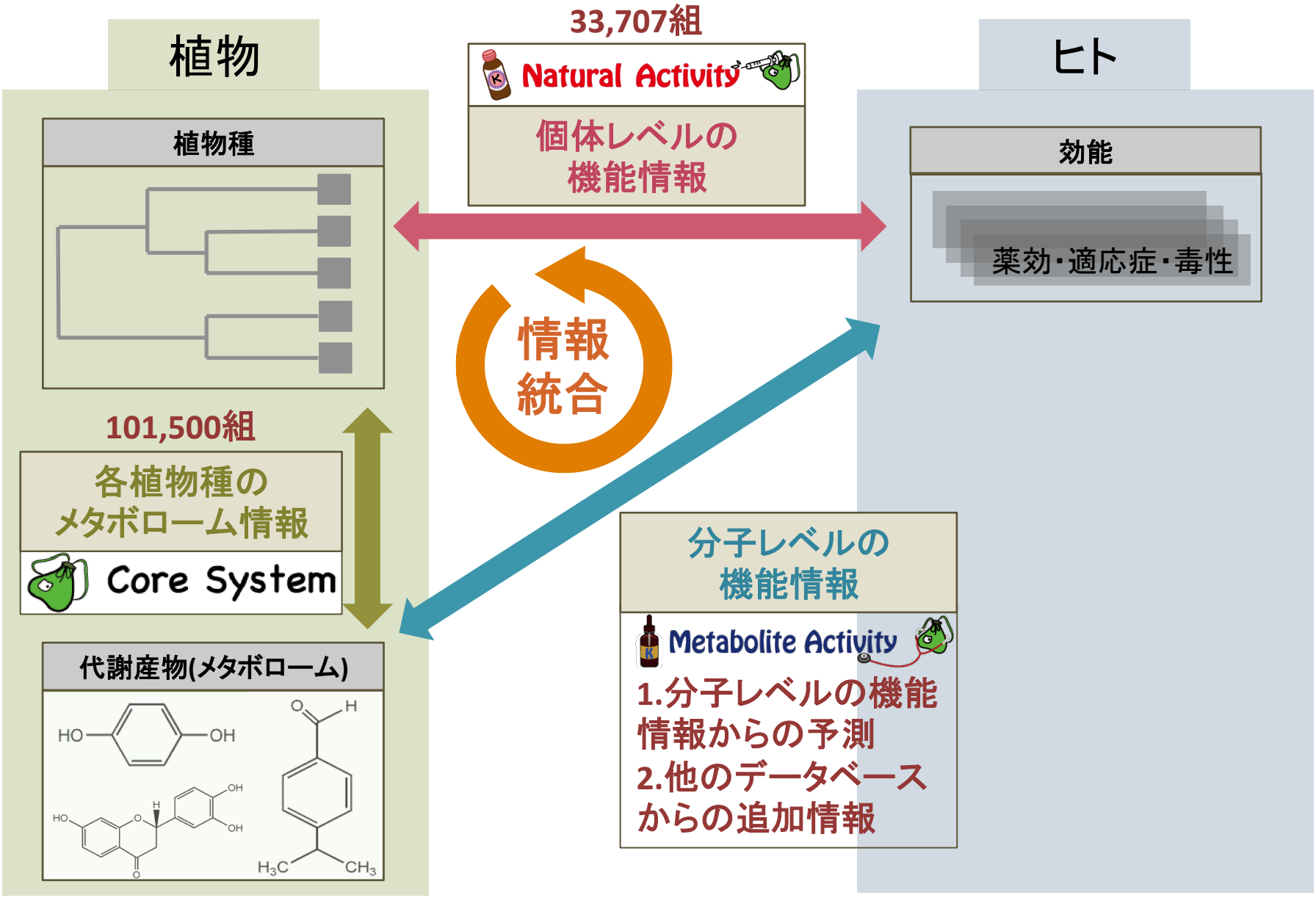
KEGG

MEDICUS, ENVIRON

代謝産物(メタボローム)



個体レベルの機能情報の予測のための統合スキーム

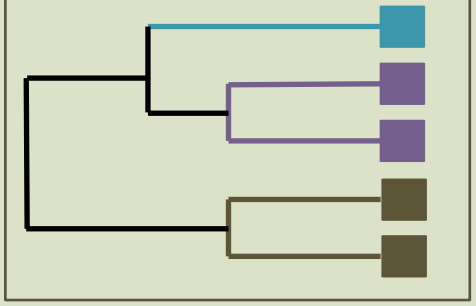


各情報のグループ化による特徴抽出

植物

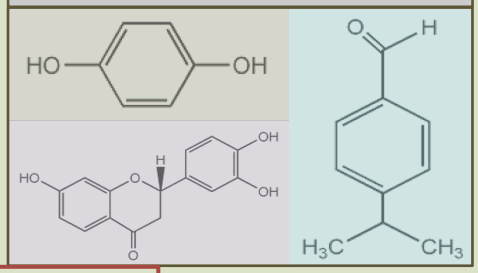
系統分類

植物種



各植物種の
メタボローム情報

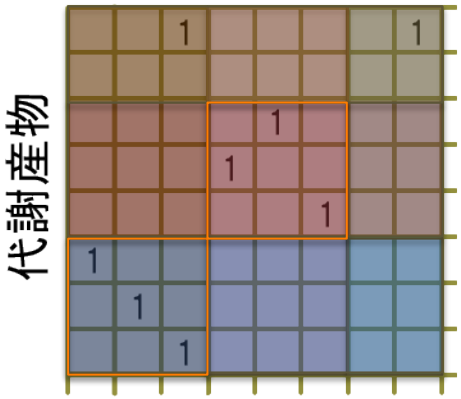
代謝産物(メタボローム)



構造分類

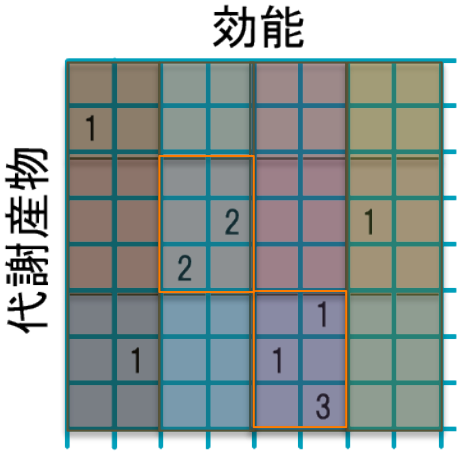
植物種

代謝産物



- 植物種のグルーピング
- 代謝産物のグルーピング

各情報のグループ化による特徴抽出



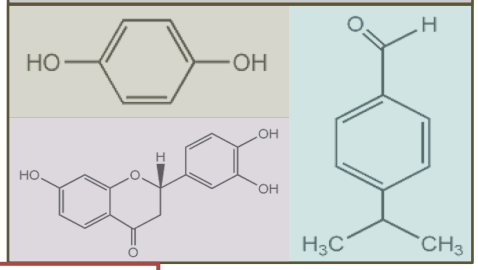
疾患分類

効能

薬効・適応症・毒性

分子レベルの
機能情報

代謝産物(メタボローム)



構造分類

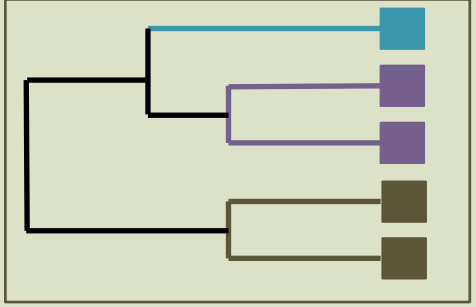
- 効能のグルーピング
- 代謝産物のグルーピング

個体レベルの機能情報の予測のための統合スキーム

植物

系統分類

植物種



ヒト

疾患分類

効能

薬効・適応症・毒性

個体レベルの
機能情報

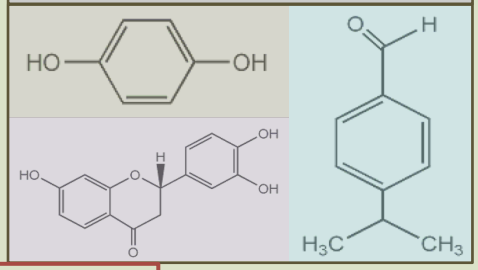
?

1.
予測基盤
サーバー

各植物種の
メタボローム情報

分子レベルの
機能情報

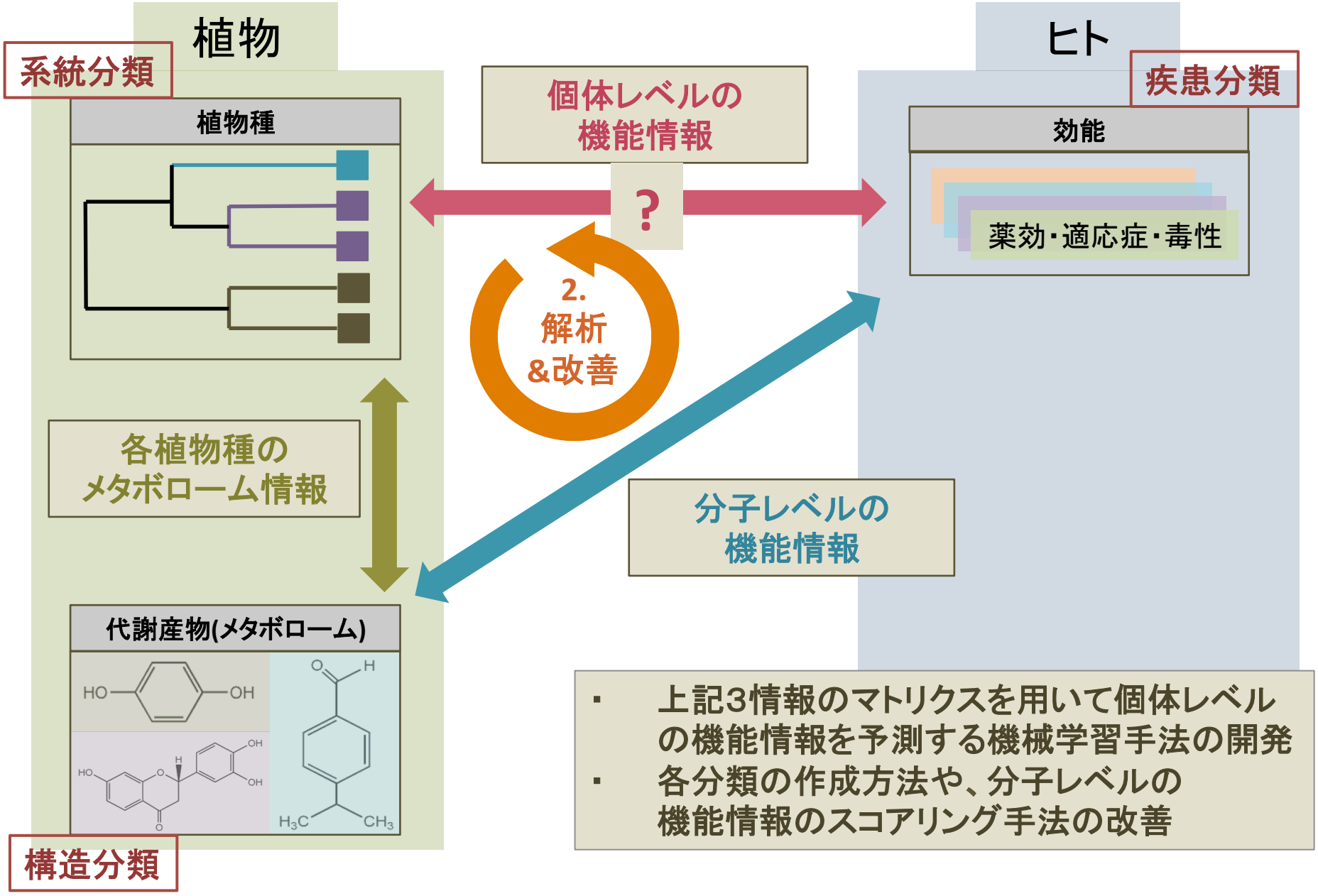
代謝産物(メタボローム)



構造分類

- 上記3情報に関して、任意の（階層）分類に対してクエリーを投げられるAPIの導入
- バッチダウンロード機能

個体レベルの機能情報の予測のための統合スキーム

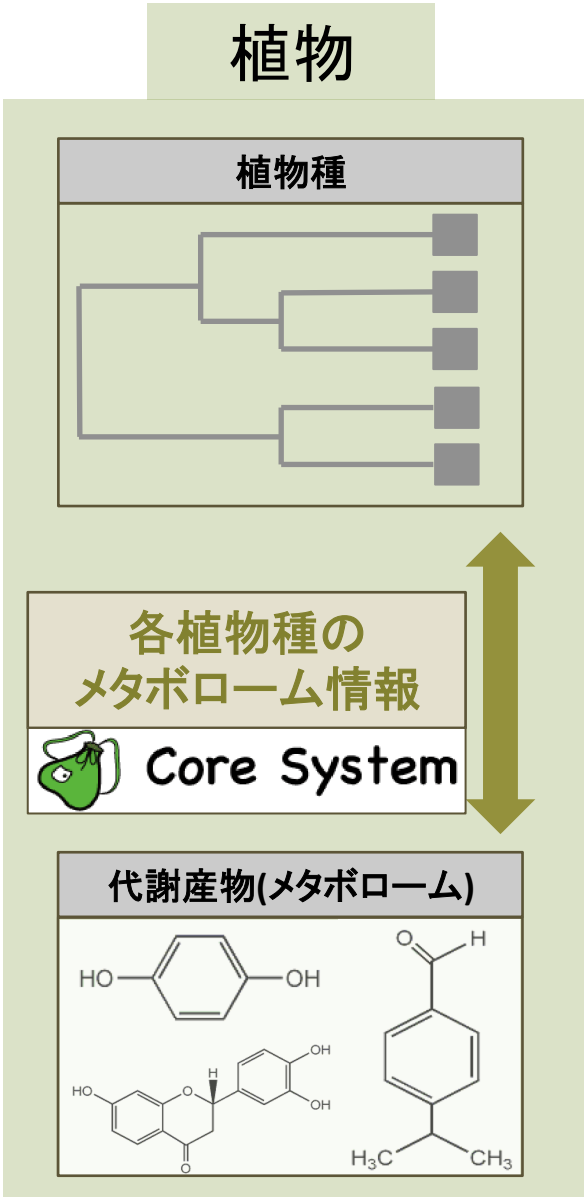


- 上記3情報のマトリクスを用いて個体レベルの機能情報を予測する機械学習手法の開発
- 各分類の作成方法や、分子レベルの機能情報のスコアリング手法の改善

目次

- 研究開発計画の概要
- KNApSAcK
  - NCBI taxonomy を用いた生物種名の標準化と階層分類
  - 代謝産物の構造分類手法
- ChEMBL、CTDとの情報統合
  - Standard InChIやCAS RNを用いた化合物リンク作成
    - 作成されたリンクの評価
- 今後の予定
  - 相互作用タンパク質のIDマッピングと分類
  - 植物種を入力して関連する疾患リストを出力
    - webインターフェイス、APIの設計

KNApSAcK coreのデータ



KNApSAcK coreのデータ

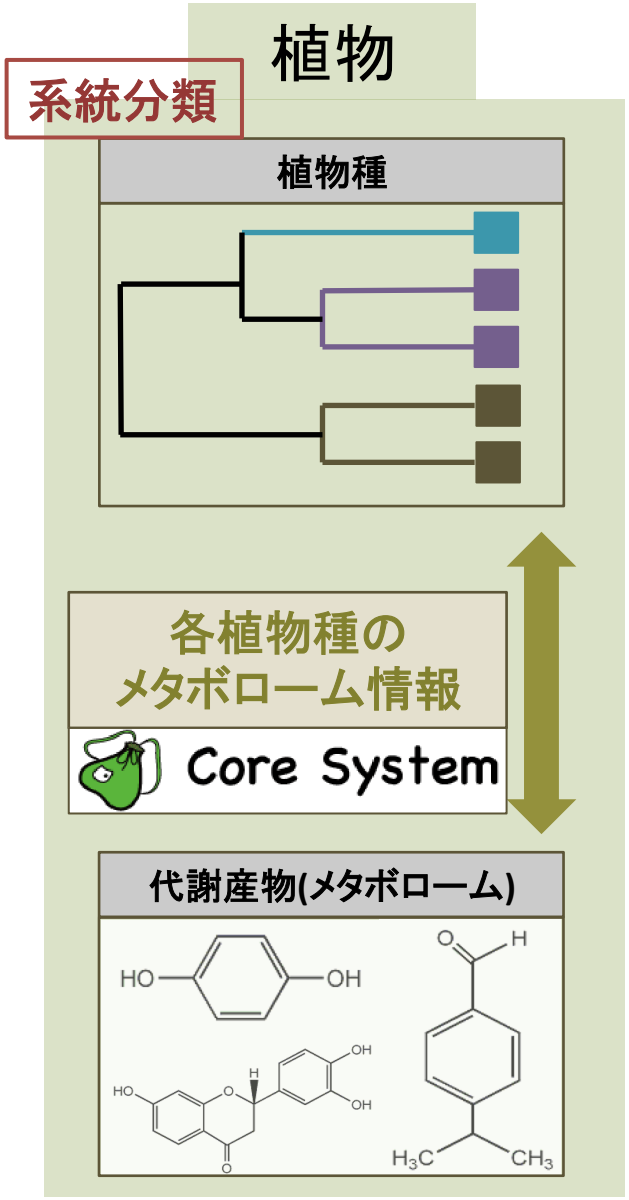
代謝産物 (50,048)　　生物種 (23,834)　　文献 (29,562)

	A	B	C	D	E	F	G	H
1	C_ID	Metabolite	Molecular Formula	Organism	Kingdom	Family	Genus	Reference
2	C00000001	Gibberellin A1;GA1	C19H24O6	Vigna unguiculata	Plantae	Fabaceae	Vigna	Garcia-Martinez,Plant Physiol.,85,(1987),212
3	C00000001	Gibberellin A1;GA1	C19H24O6	Vitis vinifera	Plantae	Vitaceae	Vitis	Perez,Am.J.Viticulture,51,(2000),315
4	C00000001	Gibberellin A1;GA1	C19H24O6	Zea mays	Plantae	Poaceae	Zea	Fujioka,Proc. Natl. Acad. Sci. USA,85,(1988),9031
5	C00000001	Gibberellin A1;GA1	C19H24O6	Bacillus licheniformis	Bacteria	Bacillaceae	Bacillus	Guitierri-Manero,Physiol.Plant,111,(2001),206
6	C00000001	Gibberellin A1;GA1	C19H24O6	Bacillus pumilus	Bacteria	Bacillaceae	Bacillus	Guitierri-Manero,Physiol.Plant,111,(2001),206
7	C00000001	Gibberellin A1;GA1	C19H24O6	Begonia x cheimantha	Plantae	Begoniaceae	Begonia	Oden,Physiol. Plant.,73,(1988),445
8	C00000001	Gibberellin A1;GA1	C19H24O6	Alstroemeria hybrida	Plantae	Alstroemeriaceae	Alstroemeria	Kappers,J.Plant Growth Regul.,16,(1997),219
9	C00000001	Gibberellin A1;GA1	C19H24O6	Althaea rosea	Plantae	Malvaceae	Althaea	Harada,Phytochem.,6,(1967),1695
10	C00000001	Gibberellin A1;GA1	C19H24O6	Arabidopsis thaliana	Plantae	Cruciferae	Arabidopsis	Talon,Heynh. Planta,182,(1990),501
11	C00000001	Gibberellin A1;GA1	C19H24O6	Aralia cordata	Plantae	Araliaceae	Aralia	Nishijima,Biosci. Biotech. Biochem.,57,(1993),1953
12	C00000001	Gibberellin A1;GA1	C19H24O6	Archidendron microcarpu	Plantae	Fabaceae	Archidendron	Koshioka,Agric. Biol. Chem.,50,(1986),1899

ftp://ftp.biosciencedbc.jp/archive/knapsack/LATEST/knapsack_core.zip

代謝産物と生物種のペア数: 101,443

KNApSAcK生物種をNCBI taxonomyにマッピング



NCBI taxonomyを利用し、生物種名を標準化
NCBIの生物種階層分類から、代謝産物が検索可能に

KNApSAcK生物種をNCBI taxonomyにマッピング

生物種 ⇛ NCBI taxonomy にマッピング

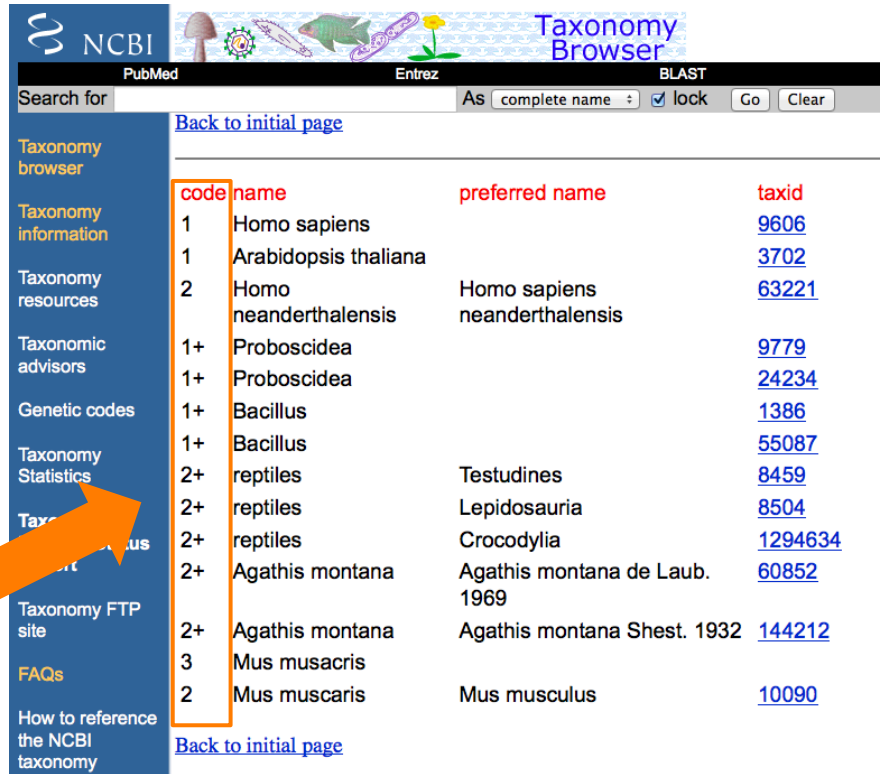
	A	B	C	D	E	F	G	H
1	C_ID	Metabolite	Molecular Formula	Organism	Kingdom	Family	Genus	Reference
2	C00000001	Gibberellin A1;GA1	C19H24O6	Vigna unguiculata	Plantae	Fabaceae	Vigna	Garcia-Martinez,Plant Physiol.,85,(1987),212
3	C00000001	Gibberellin A1;GA1	C19H24O6	Vitis vinifera	Plantae	Vitaceae	Vitis	Perez,Am.J.Viticulture,51,(2000),315
4	C00000001	Gibberellin A1;GA1	C19H24O6	Zea mays	Plantae	Poaceae	Zea	Fujioka,Proc. Natl. Acad. Sci. USA,85,(1988),9031
5	C00000001	Gibberellin A1;GA1	C19H24O6	Bacillus licheniformis	Bacteria	Bacillaceae	Bacillus	Guitierri-Manero,Physiol.Plant,111,(2001),206
6	C00000001	Gibberellin A1;GA1	C19H24O6	Bacillus pumilus	Bacteria	Bacillaceae	Bacillus	Guitierri-Manero,Physiol.Plant,111,(2001),206
7	C00000001	Gibberellin A1;GA1	C19H24O6	Begonia x cheimantha	Plantae	Begoniaceae	Begonia	Oden,Physiol. Plant.,73,(1988),445
8	C00000001	Gibberellin A1;GA1	C19H24O6	Alstroemeria hybrida	Plantae	Alstroemeriaceae	Alstroemeria	Kappers,J.Plant Growth Regul.,16,(1997),219
9	C00000001	Gibberellin A1;GA1	C19H24O6	Althaea rosea	Plantae	Malvaceae	Althaea	Harada,Phytochem.,6,(1967),1695
10	C00000001	Gibberellin A1;GA1	C19H24O6	Arabidopsis thaliana	Plantae	Cruciferae	Arabidopsis	Talon,Heynh. Planta,182,(1990),501
11	C00000001	Gibberellin A1;GA1	C19H24O6	Aralia cordata	Plantae	Araliaceae	Aralia	Nishijima,Biosci. Biotech. Biochem.,57,(1993),1953
12	C00000001	Gibberellin A1;GA1	C19H24O6	Archidendron microcarpu	Plantae	Fabaceae	Archidendron	Koshioka,Agric. Biol. Chem.,50,(1986),1899

ftp://ftp.biosciencedbc.jp/archive/knapsack/LATEST/knapsack_core.zip

Organism, Genus, Family, Kingdom をすべてクエリーとして使用し、
マッピングできたもののうち最も下の階層のものを採用

NCBI taxonomy

NCBI taxonomyのcgiサイトを利用
http://www.ncbi.nlm.nih.gov/Taxonomy/TaxIdentifier/tax_identifier.cgi

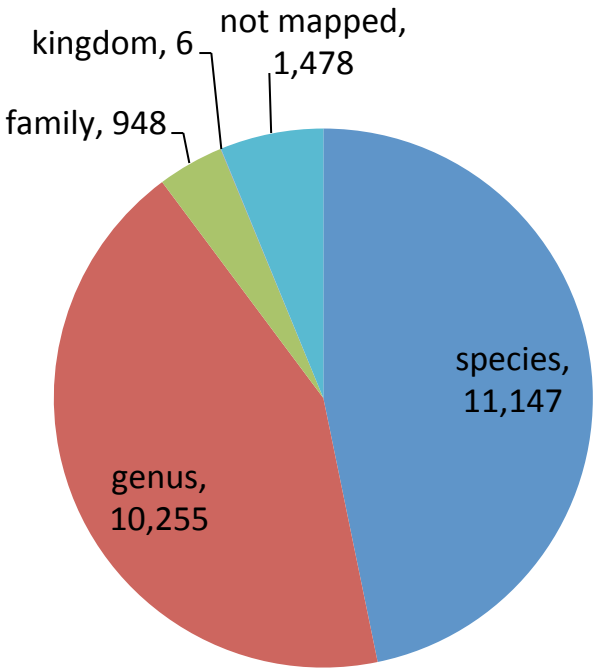


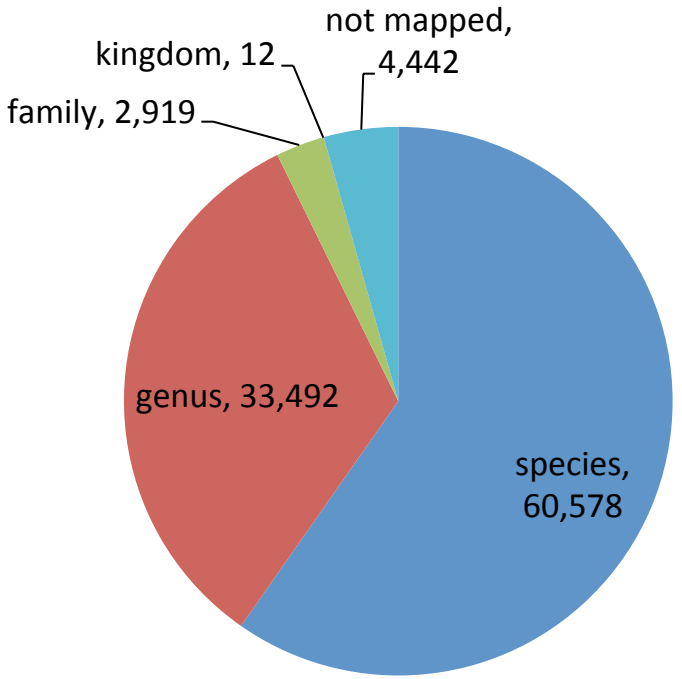
Status Codeの説明

1 - the incoming name is our primary name for a taxon in our database
2 - the incoming name is a secondary name for a taxon in our database (it could be listed as a synonym, a misspelling, a common name, or several other nametypes)
3 - the incoming name is not found in our database
+ - the incoming name is duplicated in our database (used in combination with the other status codes)

1, 2でマッピング成功

マッピング結果(マッピングレベル)



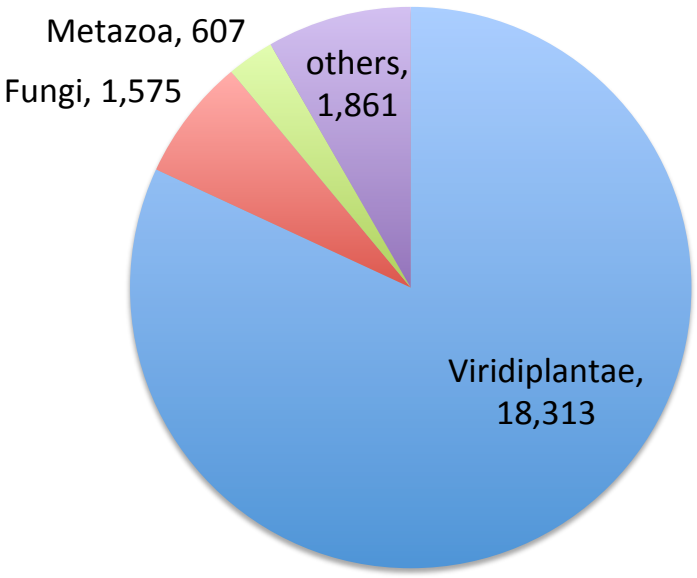


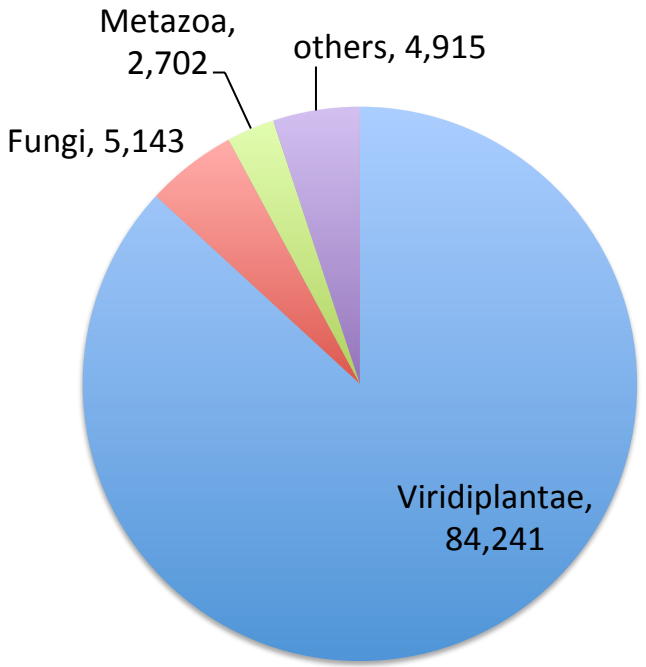
#生物種
speciesとgenusレベルで90%。

#代謝産物-生物種ペア
speciesとgenusレベルで93%。

KNApSAcKはNCBI taxonomyに概ねgenusレベル以下でマッピング可能である

マッピング結果(NCBIのKingdom分類)



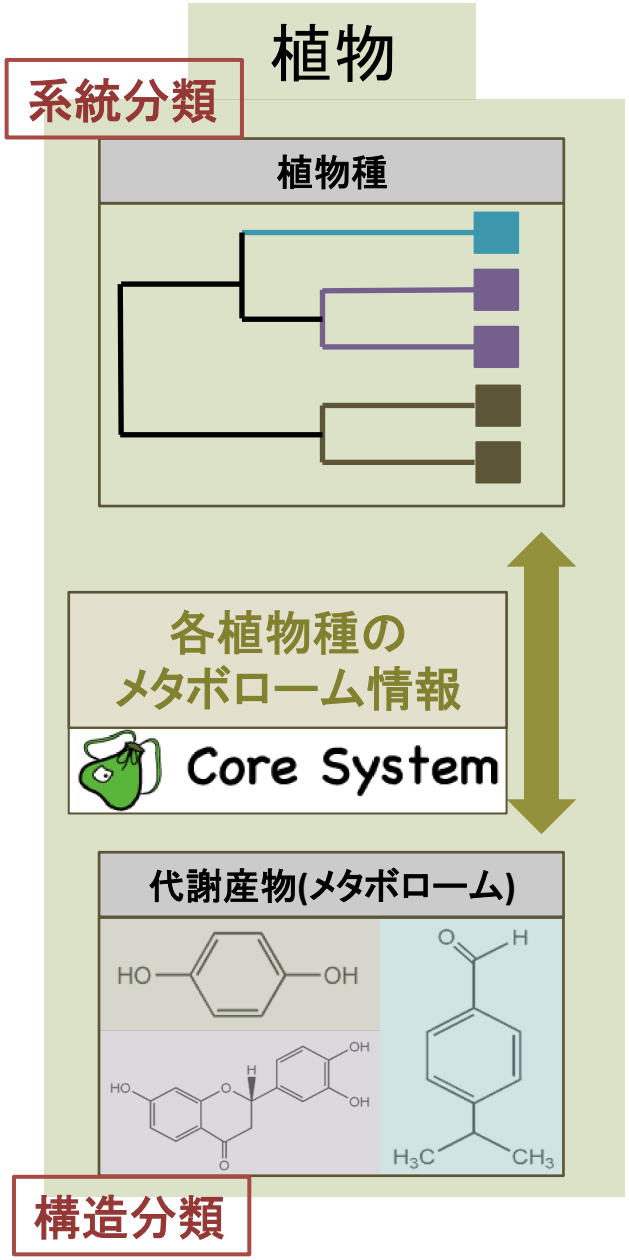


#生物種
植物: 82%

#代謝産物-生物種ペア
植物: 87%

KNApSAcKは大部分が植物に関するデータであることを確認

代謝産物の構造分類手法(開発中)



- 部分構造を利用した分類
  - PubChem fingerprint
    - 881種類の部分構造パターンによる表現(binary vector)
  - KCF-S (GIW2013, Kotera *et al.*,)
    - KEGG Atom Typeを用いた部分構造表現(integer vector)
- 母核を用いた化合物分類(予定)
  - KEGG BRITEの分類を利用
    - br:08003, Phytochemical Compounds

目次

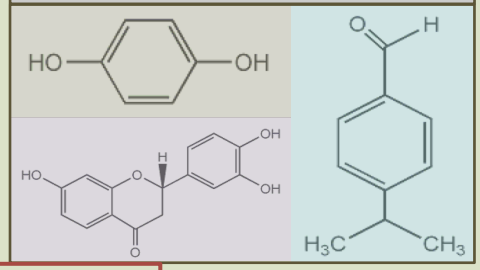
- 研究開発計画の概要
- KNApSAcK
  - NCBI taxonomy を用いた生物種名の標準化と階層分類
  - 代謝産物の構造分類手法
- ChEMBL、CTDとの情報統合
  - Standard InChIやCAS RNを用いた化合物リンク作成
    - 作成されたリンクの評価
- 今後の予定
  - 相互作用タンパク質のIDマッピングと分類
  - 植物種を入力して関連する疾患リストを出力
    - webインターフェイス、APIの設計

相互作用タンパク質情報の統合

植物

ヒト

代謝産物(メタボローム)

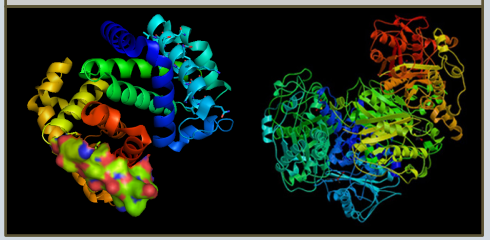


機能情報(代謝産物-タンパク)

ChEMBL

ctd™

相互作用タンパク質



構造分類

各代謝産物がもつ効能情報の統合

植物

ヒト

効能

薬効・適応症・毒性

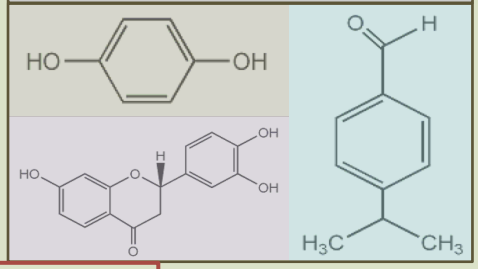
機能情報(代謝産物-効能)

ctd

KEGG

MEDICUS, ENVIRON

代謝産物(メタボローム)

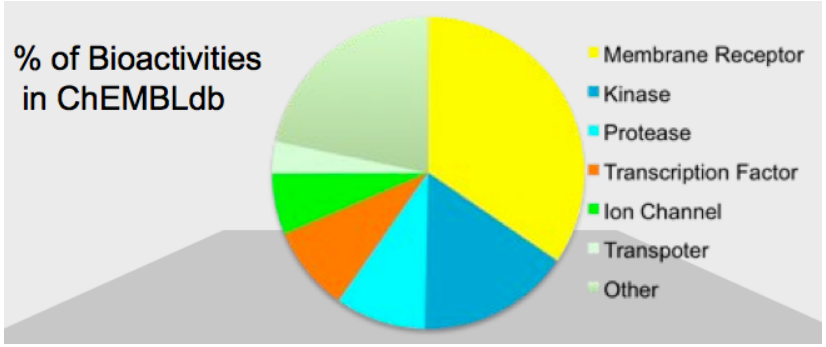


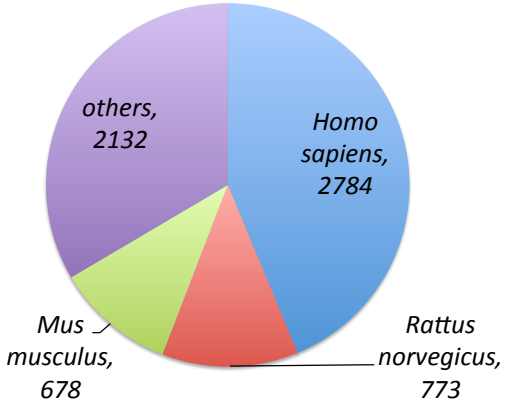
構造分類

ChEMBL

Release 17 (Aug. 2013)

- 医薬品及び開発化合物のデータベース
- ターゲット(タンパク質等)に対するアッセイ情報
  - 文献からマニュアル抽出 + PubChem からの情報が大部分を占める
  - 専門家によるキュレーション
- >73万アッセイ、>1200万活性情報、>130万化合物
- 9,356 ターゲット (e.g. “protein complex”, “cell line”, “tissue”, “organism”,,,)
- 6,367 ターゲットタンパク質 (unique UniProt ID)
  - ヒトでは2784個





タンパク質の生物種構成

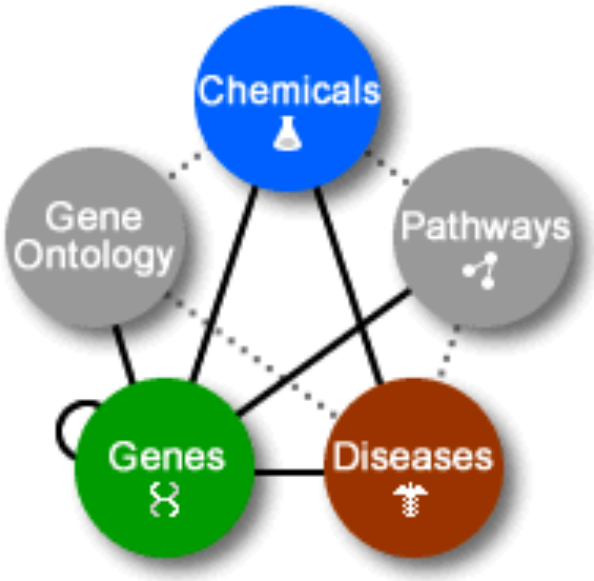
ref: http://cbi-society.org/home/documents/seminar/2009to12/CBI_Ikeda_511_d.pdf

Ver. Oct. 2013

- 環境物質が人体に与える影響を文献から抽出
- C (化合物) – G (遺伝子) 相互作用情報
  - 直接の相互作用及び、結果としてのリン酸化などの間接的な作用を含む
  - 53個の相互作用述語表現
  - C (MeSH Term), G (Entrez gene id)
  - >88万レコード (curated)

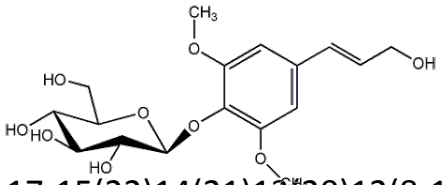
- C (化合物) – D (疾患) 相互作用情報
  - therapeutic か marker/mechanism に分類される
  - D (MeSH Term)
  - >18万レコード (curated)

- G – D 間相互作用情報
  - 文献及びOMIM由来
  - therapeutic か marker/mechanism
  - >2万8千レコード (curated)



KNApSAcKとChEMBL, CTDとの化合物リンク構築

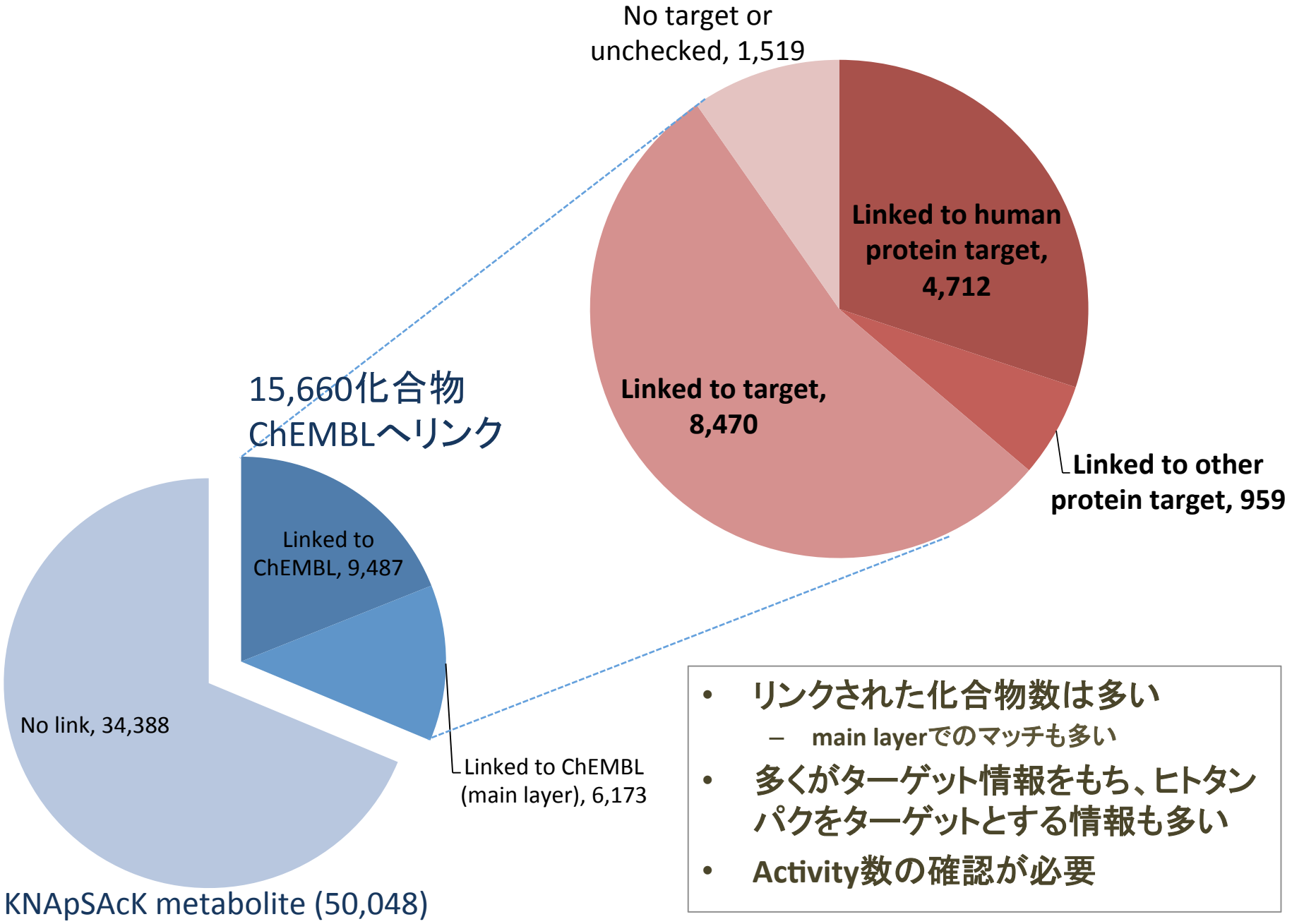
- Standard InChI
  - IUPACによる、標準的な分子構造記述法
  - 立体異性体は区別されるが、生成元のファイル(mol等)によって立体の情報量が異なることがあり、完全一致では上手くいかない事がある
  - 例) Syringin



    - KNApSAcK => “InChI=1S/C17H24O9/c1-23-10-6-9(4-3-5-18)7-11(24-2)16(10)26-17-15(22)14(21)13(20)12(8-19)25-17/h3-4,6-7,12-15,17-22H,5,8H2,1-2H3/b4-3+/t12?,13-,14+,15?,17+/m1/s1”
    - ChEMBL => "InChI=1S/C17H24O9/c1-23-10-6-9(4-3-5-18)7-11(24-2)16(10)26-17-15(22)14(21)13(20)12(8-19)25-17/h3-4,6-7,12-15,17-22H,5,8H2,1-2H3/b4-3+/t12-,13-,14+,15-,17+/m1/s1”
  - 完全一致とMain Layerでの一致の両方で進める
    - Main Layer => “InChI=1S/{formula}/c{connections}/h{H_atoms}”

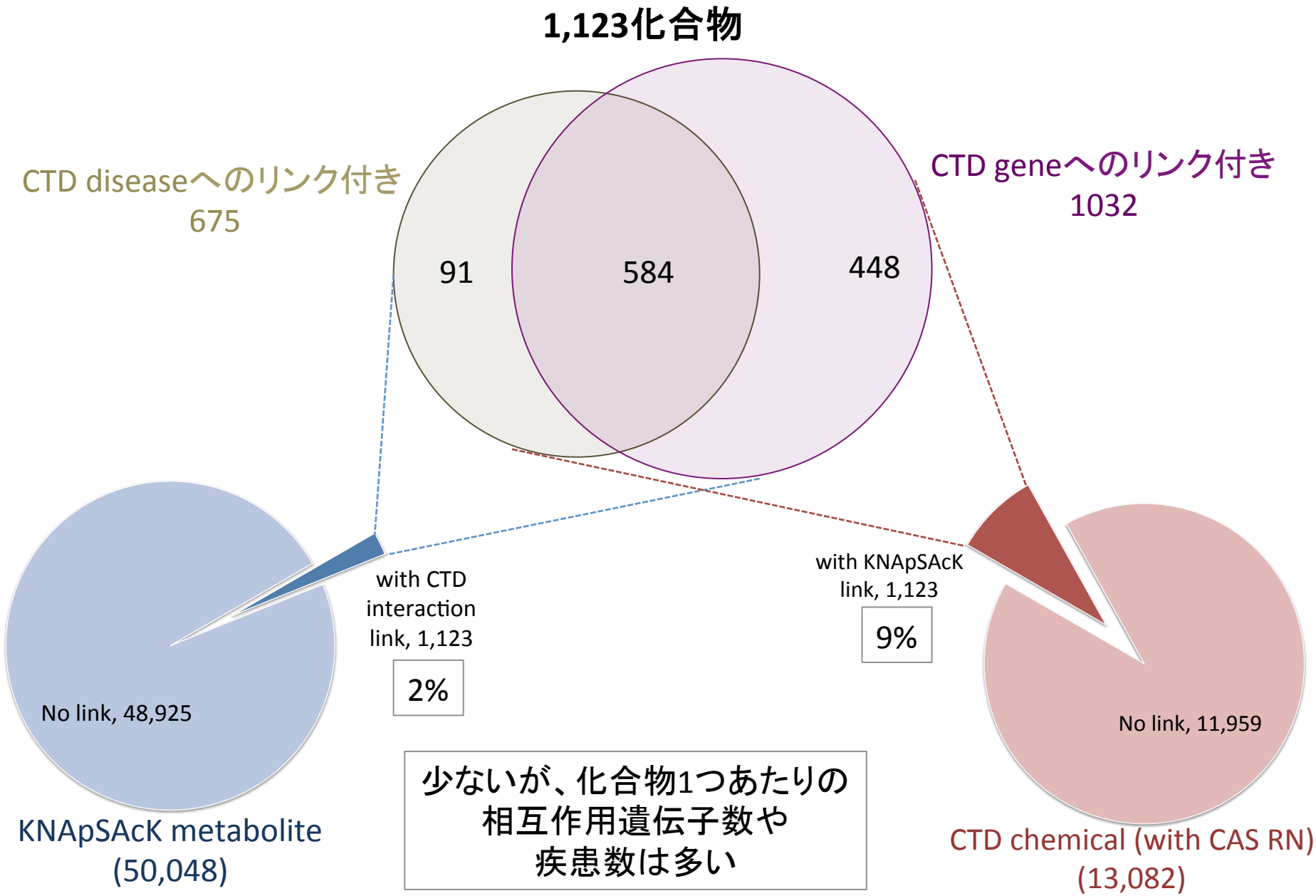
- CAS RN
  - 化学物質IDのデファクトスタンダード
  - 立体異性体は別の番号が割り当てられる

KNApSAcKとChEMBLの化合物リンク (Standard InChI)



- リンクされた化合物数は多い
  - main layerでのマッチも多い
- 多くがターゲット情報をもち、ヒトタンパクをターゲットとする情報も多い
- Activity数の確認が必要

KNApSAcKとCTDの化合物リンク (CAS RN)



目次

- 研究開発計画の概要
- KNApSAcK
  - NCBI taxonomy を用いた生物種名の標準化と階層分類
  - 代謝産物の構造分類手法
- ChEMBL、CTDとの情報統合
  - Standard InChIやCAS RNを用いた化合物リンク作成
    - 作成されたリンクの評価
- 今後の予定
  - 相互作用タンパク質のIDマッピングと分類
  - 植物種を入力して関連する疾患リストを出力
    - webインターフェイス、APIの設計

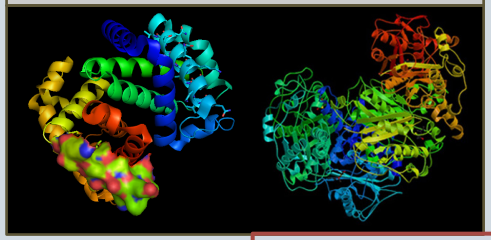
相互作用タンパク質の機能階層分類

植物

ヒト

- ID変換
  - ChEMBL: UniProt ID
  - CTD: Entrez gene ID
- 機能階層分類
  - ChEMBLの階層分類
  - Gene Ontology
  - KEGG BRITE
- KEGG BRITE

相互作用タンパク質



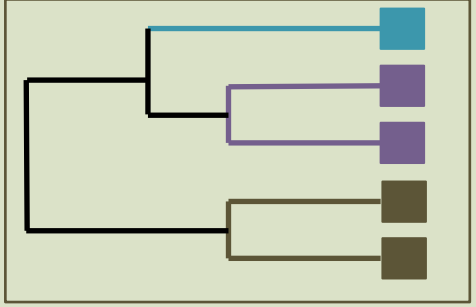
機能階層分類

各代謝産物がもつ効能情報の統合

植物

系統分類

植物種



機能情報(植物種-効能)

Natural Activity

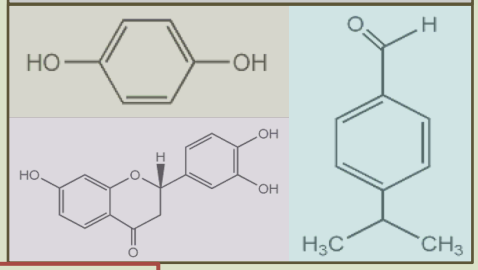
1

ヒト

効能

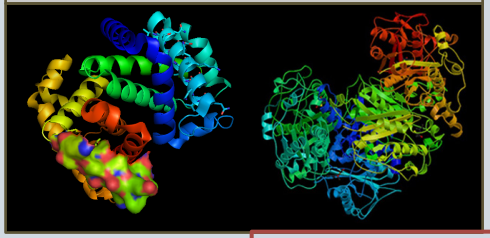
薬効・適応症・毒性

代謝産物(メタボローム)



構造分類

相互作用タンパク質



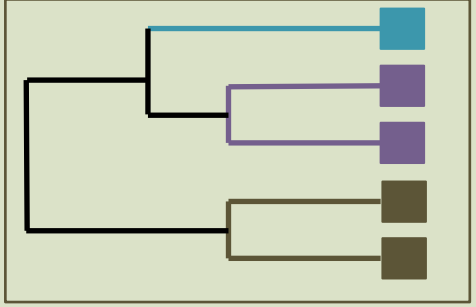
機能階層分類

各代謝産物がもつ効能情報の統合

植物

系統分類

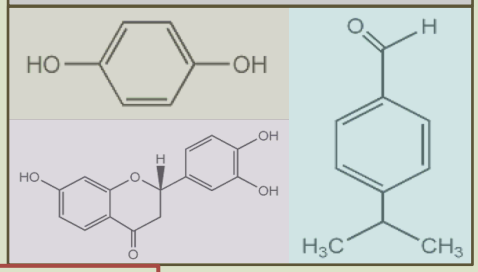
植物種



各植物種の
メタボローム情報

Core System

代謝産物(メタボローム)



構造分類

2

機能情報(代謝産物-効能)

ctd

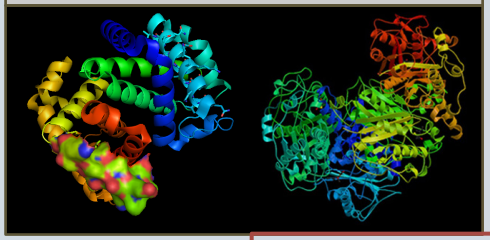
MEDICUS, ENVIRON

ヒト

効能

薬効・適応症・毒性

相互作用タンパク質



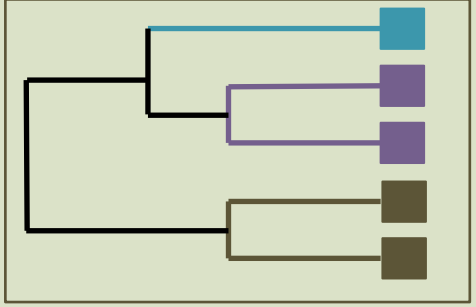
機能階層分類

各代謝産物がもつ効能情報の統合

植物

系統分類

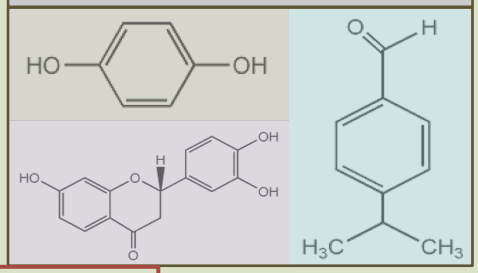
植物種



各植物種の
メタボローム情報

Core System

代謝産物(メタボローム)



構造分類

3

機能情報(代謝産物-タンパク)

ChEMBL

ctd

ヒト

効能

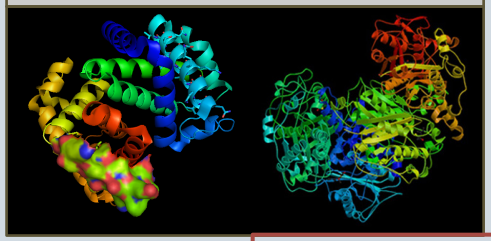
薬効・適応症・毒性

タンパク質機能情報

GO
the Gene Ontology

KEGG
PATHWAY

相互作用タンパク質



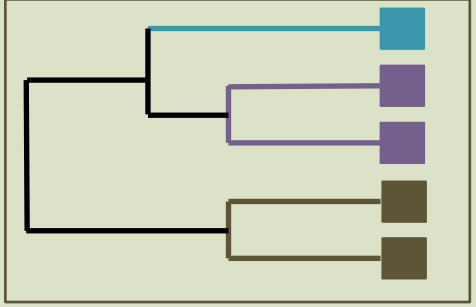
機能階層分類

各代謝産物がもつ効能情報の統合

植物

系統分類

植物種



機能情報(植物種-効能)

Natural Activity

ヒト

効能

薬効・適応症・毒性

機能情報(代謝産物-効能)

ctd

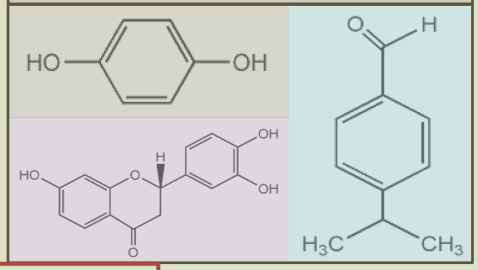
MEDICUS, ENVIRON

タンパク質機能情報

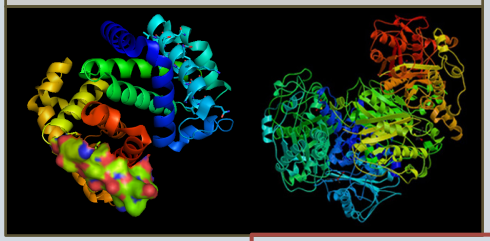
GO the Gene Ontology

KEGG PATHWAY

代謝産物(メタボローム)



相互作用タンパク質



構造分類

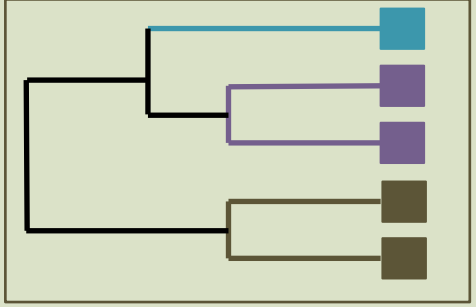
機能階層分類

各代謝産物がもつ効能情報の統合

植物

系統分類

植物種



機能情報(植物種-効能)

Natural Activity

ヒト

効能

薬効・適応症・毒性

機能情報(代謝産物-効能)

c t d

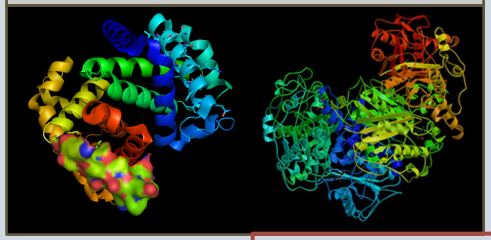
MEDICUS, ENVIRON

タンパク質機能情報

GO the Gene Ontology

KEGG PATHWAY

相互作用タンパク質



機能階層分類

疾患表現のリスト表記

1. Natural Activity
2. MeSH Term, ICD-10
   - CTD, KEGG MEDICUS
3. PATHWAY, GO
   - interacting gene enrichment

まとめ

- KNApSAcK coreの情報整理、標準化
  - 生物種分類と代謝産物構造分類
- ChEMBL、CTDの情報を統合
  - Standard InChI等による化合物リンクの作成
  - 個々のActivityについての確認を行う
- KEGGの情報を統合、タンパク質分類整備
  - GO, PATHWAY, ChEMBLの階層分類などを利用する
- 植物種と関連する疾患の予測に向けて
  - 統合した情報の公開、API等の整備