Vol.2011-BIO-25 No.33
2011/6/24

利用者の意図を考慮した概念的観点に基づく蛋白質構造解析文献検索手法

麻 生 知 希[†1]　大 川 剛 直[†1]

近年，多数の蛋白質立体構造が解析され，その結果が文献として報告されている．これに伴い，生物学・医学研究者が自分の興味の持つ文献を的確に検索するシステムに対する需要が高まっている．一方，蛋白質構造解析文献においては，用語の曖昧性や不統一性のために，単純なキーワード検索では、十分な検索精度が得られないことが問題になる．そこで本研究では複数の関連データベースを用いることで，文献と構造や機能が意味する概念を関連付け，この概念をもとに文献検索を行う仕組みを提案する．また，概念間の関連度を検索時に調整することにより，利用者の意図を反映した検索を実現する．

Method for retrieval of articles on protein structure analysis by semantic view considering user's intention

Tomoki Aso[†1] and Takenao Ohkawa[†1]

In recent years, information about protein structure and function is described in a large amount of articles. However, keyword retrieval often fails to find desired articles, because these articles contain ambiguous and complicated words. In this paper, we propose a method for retrieving articles based on semantic semilarity and user's intention.

1. はじめに

蛋白質の構造や機能の情報は医学・生物学研究者らによって日々研究され，得られた知見は蛋白質構造解析文献として発表されている．これまでに公開・蓄積されている蛋白質構造解析に関する論文の件数は約 3 万件に及び，またその登録件数は加速度的に増加していることから，これらの論文を簡単に検索できるシステムに対する要求が高まっている．

文献検索を実現する簡単な方法の 1 つとして要旨や本文からのテキスト全文検索等が考えられる．しかし，医学・生物学において論文に記載される蛋白質名や機能を表す語は統一されておらず，1 つの蛋白質に複数の別名がついていたり，同一の概念を指していても異なる語を用いている場合があることから，検索に十分な精度を得られるとは考え難い．さらに，テキスト全文検索では機能などの概念的な観点からクエリとの関連性を評価できず，研究者の意図する結果が得られるとは限らない．

蛋白質構造解析文献には様々な情報が付与されており，様々な角度から概念やその概念間の関係を評価することができる．このことから，本研究では，蛋白質構造解析文献を入力クエリとし，機能や構造の視点から類似する蛋白質構造解析文献を推薦するシステムを提案する．蛋白質構造解析文献を入力クエリにすることで，単にテキスト検索をする場合に比べ，入力クエリを概念的に捉えたり，文献間に関連する概念を特定する等の複雑な視点で評価することが可能になると考える．

しかしながらその一方で，論文そのものを入力クエリとした場合，クエリが多面的な概念を包含することになるため，研究者の意図する概念を一意に特定することが困難になり，意図に合った検索結果が得られない可能性がある．意図する結果を得るためにテキスト検索では入力語を追加して AND 検索を行うことが一般的である．そこでこれと同様に研究者が意図している観点において類似すると考える文献を付加文献として追加する枠組みを導入する．これにより研究者の意図を特定し，その意図を文献間の類似度に反映し，それに基づいた検索が可能なシステムを提案する．

2. 蛋白質構造解析文献と関連データベース

本研究では，蛋白質構造解析文献を対象とする関連文献検索を行う上で，複数のデータベースを相互に連携させて利用する．以下に本研究で利用するデータベースを示す．

- 文献データベース (MEDLINE/PubMed)
- 概念データベース (Gene Ontology)
- 蛋白質立体構造データベース (Protein Data Bank(PDB))
- 蛋白質モチーフデータベース (PROSITE)
- 蛋白質配列データベース (Swiss-Prot/UniProt)

†1 神戸大学大学院システム情報学研究科
Graduate School of System informatics, Kobe University

Vol.2011-BIO-25 No.33
2011/6/24

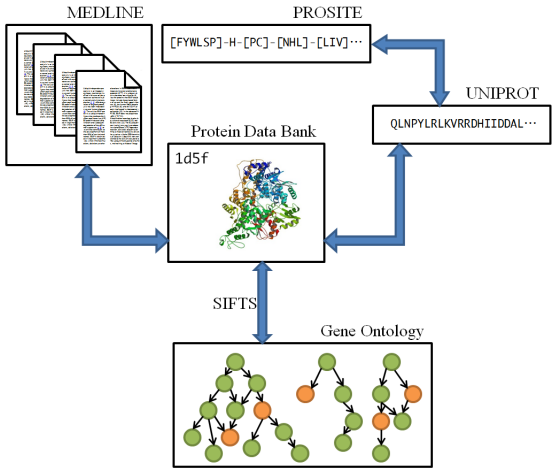

図 1 各データベース間の関係
Fig. 1 Relation between Databases.

このうち，概念データベースである Gene Ontology には蛋白質が持つ機能の名称や用語についての共通語彙である GO Term が策定されている．各 GO Term は包括関係を示す形で親子関係が定義されており，この関係から非循環有向グラフの構築ができる．

また，蛋白質立体構造データベースである Protein Data Bank(PDB) には蛋白質ごとに立体構造のデータや PDB ID 等が付与されており，SIFTS というプロジェクトによって GO Term に対応付けられている．

これらのデータベースを図 1 のように連携利用する．すなわち，MEDLINE に掲載されている蛋白質構造解析文献はその蛋白質の構造が PDB に登録されている．PDB に登録された蛋白質の概念は SIFTS によって Gene Ontology と関連付けられ，各概念は非循環有向グラフ上にマッピングされる．また，蛋白質の配列をもとに UniProt を利用することで，類似する構造を持つ蛋白質の情報を得ることができる．

3. 概念関係データベースを用いた概念類似度の算出

3.1 概　　要

蛋白質構造解析文献には概念や蛋白質の特性などを示すのに複雑な専門用語が用いられており，統一する規格も存在しないため，同じ概念を表現するのに同一の記述がなされているとは限らない．そこで，本研究では Gene Ontology を利用し，蛋白質の構造や機能を統一された概念を用いて捉える．そして，入力文献と検索対象文献間の類似度を算出する必要がある．これは，入力文献と検索対象文献のそれぞれの文献に付与された概念を用いて，その概念間の類似度を算出することで文献間の類似度とする．

さらに，入力クエリを単一ではなく複数とし，その複数の入力クエリがどのような概念について関連しているかを特定して，これをユーザの意図として検索結果に反映させることを考える．具体的には，関連性を有する概念間は，そのユーザの意図において類似していると見なせるように類似度を補正することで，同様の観点からの概念を持つ文献が検索されやすいようにする．

3.2 Gene Ontology 非循環有向グラフ上における最短経路の定義

概念類似度を算出する際に用いるために，Gene Ontology の非循環有向グラフ上の離れた位置にある 2 つの概念 (グラフ上では 2 つのノード) の最短距離で結ぶ経路をグラフ上での最短経路として定義する．Gene Ontology の非循環有向グラフの特徴として，根に近づくごとに一般的な語となり，葉に近づくごとにより細分化された語となる．この特徴から，単に最短経路をとることが無意味になる場合がある．例えば，cellular nitrogen compound metabolic process(GO:0034641) と cellular amide metabolic process(GO:0043603) の 2 概念間の経路長を算出する例を図 2 に示す．この 2 概念間の最短経路は，図 2(a) のように両者の下位概念である biotin metabolic process(GO:0006768) を経由した経路であり，その経路長は各々の弧の長さを 1 とすると 2 となり，最短である．

しかし，Gene Ontology の概念間の関係は “下位の概念は上位の概念より具体的にした概念であり，上位の概念は下位の概念をより一般的な概念である” と定義されている．そのため，図 2(a) のように下位の概念を経由した経路を最短経路とするのはこの定義に沿わないと考えられる．

このことから，ここでは Gene Ontology 上のノード間の最短経路を次のように定義する[1)]．まず，2 つの概念の共通の上位概念となるノードを探す．この共通の上位概念とそれぞれの概念間の経路長の和が最小となる経路を探し，これを繋げたものを最短経路とする．これを適用したものが図 2(b) である．この定義に従えば，最短経路は共通の上位概念である cellular metabolic process(GO:0044237) を通る経路 (経路長は 3) である．

3.3 概念関係を用いた類似度算出

3.3.1 概念間の最短経路長

概念類似度を算出するにあたり，3.2 節で定義した最短経路を利用する．この最短経路長

Vol.2011-BIO-25 No.33
2011/6/24

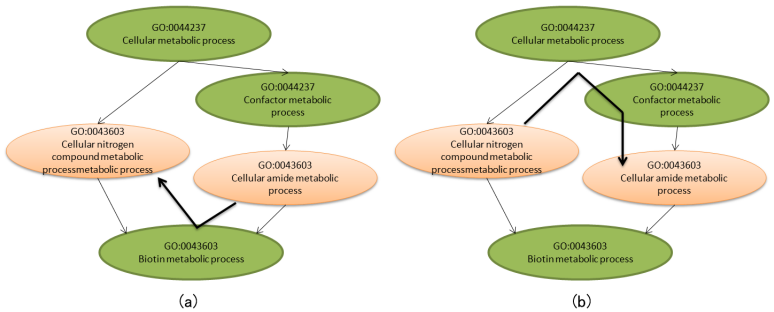

図 2 GO:0034641，GO:0043603 間の最短経路
Fig. 2 Shortest path between two GO Terms,GO:0034641 and GO:0043603

が短い概念ほどより類似した概念と考えることができる．最短経路を構成する弧 p の集合を P とすると，2 概念 (t_1, t_2) 間の最短経路長 $len(t_1, t_2)$ を次式のように定義する．

$$len(t_1, t_2) = \sum_{p \in P} w(p) \tag{1}$$

ここで，$w(p)$ は弧 p に付与された重みである．特に明記しない場合は，$w(p) = 1$ である．

3.3.2 概念間の類似度

2 概念 (t_1, t_2) 間の類似度 $Sim^{GO}(t_1, t_2)$ は最短経路長 $len(t_1, t_2)$ を用いて定義する．ここでは概念間の類似度算出には，概念間の距離が短いほど類似度が高くなる特徴があり，正規化を考慮した *normilized path length*[2)3)] を用いて以下のように定義する．

$$Sim^{GO}(t_1, t_2) = \begin{cases} 1 - \frac{log(len(t_1, t_2))}{log(2 * maxdepth^{GO})} & \text{if } t_2 \in c(t_1) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

ここで，$c(t)$ は概念 t を含むグラフの全てのノードである．$maxdepth^{GO}$ は根から最も遠い概念までの経路長である．現在の Gene Ontology のグラフで最も根から遠い概念までの経路長は 12 であるので，$maxdepth^{GO} = 12$ となる．

この式を用いることによって経路長を類似度に変換することができる．また，$maxdepth^{GO}$ によって正規化がなされており，最も類似する概念の類似度 (すなわち類似度を計算する 2 概念が同一の概念であった場合の類似度) は 1，遠ざかるごとに 0 に近づくこととなる．

3.3.3 概念集合間の類似度算出

一般に 1 つの文献には複数の概念が付与されており，このように複数の概念から構成される概念集合間で類似度を算出する場合，概念ごとに類似するかどうかを調べ，それぞれの類似度に高いものが多いならば概念集合間についても類似度が高いということができる．このことを利用して，概念集合 T_1 の概念のそれぞれについて，別の概念集合 T_2 のうち最も類似度の高い概念を注目し，それらの類似度を平均して算出する．定式化すると次のようになる．ここで，$|T|$ は概念集合に含まれる概念数を示す．

$$Sim^{GOs}(T_1, T_2) = \frac{\sum_{t_1 \in T_1} \max_{t_2 \in T_2} Sim^{GO}(t_1, t_2)}{|T_1|} \tag{3}$$

ここで注意したいのは，この式に主従関係があることである．すなわち，(3) 式では T_1 のそれぞれの概念について概念類似度が最大となる概念を算出し，その平均をとっている．すなわち，T_2 は全ての概念が用いられるのではなく，類似度計算において T_2 のどの概念が利用されるかは，T_1 の概念に依存することになる．このように T_1 を一意にすることで，概念数 (ここでは $|T_1|$) 等を統一して様々な概念集合との比較が可能となる．

3.3.4 ユーザの意図の特定

ユーザが検索を行う際には，何らかの意図が働いている．このとき，その意図のもとで，どのような概念関係が重視されているかを特定し，これを検索結果に反映させる．具体的には，意図によって重視された概念関係を示す経路上の弧にかかる重みを 1 より小さくすることで最短経路長を短くし，ユーザが意図する概念関係を持つ文献間の類似度を高く評価する．

このために，(1) 式の $w(p)$ を以下のように計算する．まず重みを変化させる範囲であるが，検索のクエリとして入力されていた概念を t_m，意図を反映させるために追加された概念を t_r とおくと，t_m と t_r の間に注目された概念関係があると考える．t_m 及び t_r より一般的な概念は全てその概念に関連していると考えることができるので，重み付けをする範囲は弧の両端が t_m 及び t_r の全ての上位概念 (t_m 及び t_r を含む) によって構成されている全ての弧とする．重み付けの対象となる弧の中でも，t_m 及び t_r に近い弧は，より詳細な (細分化された) 関係であり，遠い弧はより一般的な関係であるから，t_m 及び t_r に近い弧はより強い重み付けを行い，遠い弧 (すなわち一般的すぎる概念関係) にはあまり重み付けを強くしないものとし，t_m 及び t_r からの距離に応じて重みに傾斜をかけるようにする．具体的な重み付けは次式のように，重みをかける弧の両端のノードから t_m または t_r までの距離を，t_m 及び t_r からの距離の影響を強くするためにそれぞれ二乗して逆数にし，両ノードの重みの平均の値を弧の重みとする．なお，分母が 0 となることを防ぐために，分母に 1 を

加算している．t_r について，最短経路上の弧 p の両端のノードが (t, t') であった場合，まず一つの t_r がある弧 p に影響を及ぼす重み $weight(p, t_r)$ を次のように定義する．

$$weight(p, t_r) = \begin{cases} \frac{\omega}{2}\left(\frac{1}{(curv * l(t_r, t))^2+1} + \frac{1}{(curv * l(t_r, t'))^2+1}\right) \\ \qquad\qquad \text{if } t, t' \in (isUp(t_r) \cup isUp(t_m)) \\ 1 \\ \qquad\qquad \text{otherwise} \end{cases} \tag{4}$$

ここで，(t, t') は弧 p の両端のノードであり，$isUp(t)$ は t の全ての上位概念の集合である．$l(t, t')$ は概念 (t, t') 間の最短経路に含まれる弧の数である．ここには t_r のみ示したが，t_m についても t_r と同様に $weight(p, t_m)$ を算出する．

ここで，ω と $curv$ はそれぞれ重みをかける程度やその傾きを定義するパラメータである．ω は重みの強さを定義している．$\omega = 0$ の時，重みが一切かからず，ω が大きくなるごとに強い重みがかかる．$curv$ は重み付けの範囲を定義しており，$curv$ が小さくなるほど，広い範囲に重みがかかる．

入力される概念は複数存在する場合がある．その場合は，その全ての概念が有効になるように検索のクエリとして入力されていた概念の集合を T_m，追加で入力された概念の集合を T_r とすると，複数の概念間の類似度 $w^*(p, T_m, T_r)$ を次のように定義する．

$$w^*(p, T_m, T_r) = \sum_{t \in T_m, T_r} weight(p, t) \tag{5}$$

また，重み $w(p)$ を 0 から 1 の範囲に収めるために，(5) 式で算出した Gene Ontology 上の全ての弧の重みの最大値を用いて次式のように正規化を行う．

$$w(p) = \frac{w^*(p, T_m, T_r)}{\max\limits_{c \in P} w^*(c, T_m, T_r)} \tag{6}$$

4. 蛋白質構造解析文献検索システムの構築と評価

4.1 蛋白質構造解析文献検索システムの概要

本研究で構築した，複数文献の入力が可能な蛋白質構造解析文献検索システムの概要を説明する．まず，検索対象文献は PDB に登録されている蛋白質を掲載している全ての論文である．この中から，配列の類似性や機能を示すキーワードを用いて検索対象文献のフィルタリングを行う．これにより，明らかに関連のない文献を除去する．次に，入力クエリである文献とその文献に付与された概念の対応を取る．入力クエリが複数ある場合は，研究者が検索したいと考える文献 (主文献) と研究者が意図を反映するために入力した文献 (付加文献) に分けたうえで，検索において着目されている意図を特定し，Gene Ontology の非循環有向グラフに重みを付ける．こうして得られた重み付き非循環有向グラフを用いて検索対象文献と入力文献間の類似度を 3 章で述べた方法で算出し，その類似度順にランク付けをする．

4.2 検索対象のフィルタリング

検索対象文献となるのは PDB に登録されている蛋白質の構造解析をした全ての論文であるが，その総数は現在 3 万件であり，今後さらに多くなるであろうと予想できる．これらの蛋白質は全て入力文献と関連するとは限らず，あまり関連の無い蛋白質との類似度を算出することは意味がない．また，これら全ての論文を検索することは検索時間を長くすることに繋がり，システムのリアルタイム性を損なうことになる．そこで無関係な論文を検索対象から外すために，検索対象文献をふるいにかけ，関係のあるものを取り出す．

具体的には，まず，PROSITE を用いる．PROSITE には配列が類似した蛋白質がファミリーとして一纏めにされている．このファミリーの情報を用いることで，配列として類似した蛋白質を得ることができる．PROSITE において入力文献と関連する蛋白質と，同一ファミリーに属する蛋白質を有する蛋白質構造解析文献を検索対象文献とする．

しかし，PROSITE で同一ファミリーに登録されている蛋白質が数個しかない場合がある．検索対象文献が数件となると精度の良い検索を保証できないので，検索の範囲を広げる必要がある．そこで，PDB に登録されているキーワード (Descriptor) を用いる．PDB には表 1 のように蛋白質に関係するキーワードが登録されている．このキーワードと同じ単語を持つ蛋白質を取り出す．例えば，PDB ID:1dgs のキーワードは “LIGASE” という単語が PDB ID:1c4z のキーワードと共通している．この例のように，入力文献に掲載されている蛋白質に付与されたキーワードと共通の単語が存在する蛋白質 (表 1 では PDB ID:1c4z が入力文献の場合，PDB ID:1yh2, PDB ID:1e0d,PDB ID:1dgs, PDB ID:2oni が該当する) はそれぞれ似た性質があるものと考え，これらも検索対象文献に加える．

このようにして PROSITE のファミリーの情報と，PDB のキーワードを利用して検索対象文献を絞り込み，精度の向上を図る．

4.3 提案システムの評価

4.3.1 蛋白質構造解析文献の検索例

入力として一文献，“Structure of an E6AP-UbcH7 complex: insights into ubiquitination by the E2-E3 enzyme cascade.(PDB ID:1c4z)” について検索した結果の上位 5 件と

表 1 PDB の descriptor
Table 1 PDB descriptor

PDB ID	PDB Descriptor
1c4z	UBIQUITIN-PROTEIN LIGASE E3A/UBIQUITIN CONJUGATING ENZYME E2
1yh2	HSPC150 protein similar to **ubiquitin-conjugating** enzyme
1e0d	UDP-N-ACETYLMURAMOYLALANINE–D-GLUTAMATE **LIGASE**
1dgs	DNA **LIGASE** FROM T. FILIFORMIS
1dhp	DIHYDRODIPICOLINATE SYNTHASE
2oni	E3 **ubiquitin-protein** ligase NEDD4-like protein (E.C.6.3.2.-)
1uby	FARNESYL DIPHOSPHATE SYNTHASE, DIMETHYLALLYL DIPHOSPHATE

その文献のタイトルを表 2 にまとめる．

表 2 PDB ID:1c4z の検索結果上位 5 件
Table 2 Search result top5 of PDB ID:1c4z

PDB ID	Title
1z5s	Insights into E3 ligase activity revealed by a SUMO-RanGAP1-Ubc9-Nup358 complex.
1fbv	Structure of a c-Cbl-UbcH7 complex: RING domain function in ubiquitin-protein ligases.
2nvu	Basis for a ubiquitin-like protein thioester switch toggling E1-E2 affinity.
2c2v	Chaperoned ubiquitylation-crystal structures of the CHIP U box E3 ubiquitin ligase and a CHIP-Ubc13-Uev1a complex.
2grn	Lysine activation and functional analysis of E2-mediated conjugation in the SUMO pathway.

この表 2 のうち，PDB ID:1z5s と PDB ID:2grn の Title に現れる SUMO とは，Small Ubiquitin-related(like) Modifier の略称であり，Ubiquitin(遺伝子等の作用に重要な役割を示す蛋白質) という概念について類似したものである．この SUMO のような語はテキスト検索によって Ubiquitin との関連を見出すのは難しい．

表 2 より，PDB ID:1c4z の Title にある，UbcH7 という蛋白質や SUMO，Ubiquitin などの概念的に類似する蛋白質構造解析文献を抽出できていることがわかる．

4.3.2 複数入力による蛋白質構造解析文献の検索

ユーザの意図を反映した検索ができているかどうかの評価を行うために，正解データを設定する．本研究では入力文献全て，すなわち主文献と付加文献を共に引用している文献 (蛋白質構造解析文献に限らない) が，他に引用している蛋白質構造解析文献を正解データとして扱う．また，結果はランキング形式で出力されるため，平均適合率 (Average Precision/AP) を用いて評価する．

さらに，複数文献の入力によって重み付けを行う際に (4) 式に従って，ω と $curv$ の二種類のパラメータを設定する必要がある．今回はこのパラメータを ω，$curv$ のそれぞれについて 0 から 1 までの 0.1 刻みで変化させて，全てのパラメータについて平均適合率を求める．これを複数の入力セットについて行う．入力セットは 3 文献，PDB ID:1c4z，PDB ID:1fbv，PDB ID:1ldk のいづれかを主文献とし，各々の文献についてそれぞれ PubMed で上で関連が高いとされている 7 文献を付加文献として選び，合計 21 種類の入力セットを作る．入力の全てのセットは表 3 の通りである．

表 3 入力セット
Table 3 Input articles

Main	Add	Main	Add	Main	Add
1c4z	1ayz	1fbv	1d5f	1ldk	1d5f
1c4z	1fbv	1fbv	1fqv	1ldk	1fbv
1c4z	1fxt	1fbv	1fxt	1ldk	1fqv
1c4z	1kps	1fbv	1ldk	1ldk	1nex
1c4z	1nd7	1fbv	1nd7	1ldk	1p22
1c4z	1u9a	1fbv	2e2c	1ldk	1u6g
1c4z	1y8q	1fbv	2esk	1ldk	1vcb

これらの入力セットをもとに検索を実行し，得られた検索結果の AP の平均である Mean Average Precision(MAP) を算出し，評価を行う．パラメータ ω による結果の違いを図 3 に示す．縦軸は MAP，横軸はパラメータ $curv$ の値である．

このグラフから，$\omega = 0.9$，$curv = 0.2$ の時，MAP は最大の値，0.5074 をとる．また，$\omega = 0$ の時は，付加文献の重みが全く考慮されない状態であり，これは主文献のみで検索した時と同値となり，この時の MAP は 0.4872 である．この二つの方法について Wilcoxon の符号付き順位検定によれば有意水準 $\alpha < 0.05$ において有意な差が得られている．これより，本手法の付加文献を用いた重み付けが意図の特定に有用であることが示される．

また，既存のシステムによる結果とも比較するために，PubMed を利用した検索と精度を比較する．PubMed には Related citations という検索インターフェースが存在し，そこでは文書のタイトル，概要，生命科学用語集である MeSH というデータベースを用いて PubMed

Vol.2011-BIO-25 No.33
2011/6/24

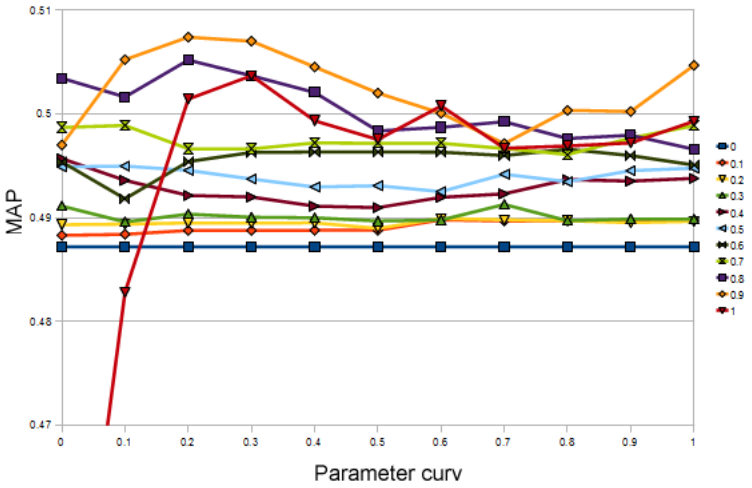

図 3 本手法により導出された結果の MAP
Fig. 3 Mean Average Precision of this system

独自の非公開のアルゴリズムで文書検索が実現されている[4]．この Related citations は蛋白質構造解析文献以外の論文も出力するが，同じ母集団での評価を行うため，蛋白質構造解析文献のみを抽出して扱う．

一方，PubMed では AND 検索を行うことができないので，以下のような手順で，主文献にも付加文献にも関連する文献が上位になるようにする．まず主文献を検索した結果のランキングの 1 位にスコア 50 を与え，2 位には 49，3 位には 48 とスコアを前の順位より 1 だけ小さくなるようにそれぞれスコアを与える．これを付加文献についても行い，主文献のスコアと付加文献のスコアを足し合わせたスコアをもとにランキング化する．この方法を表 3 の入力文献セットをそれぞれ入力して適用したところ，MAP は 0.3464 となった．この値は，図 3 で示した結果に対して明らかに下回っており，提案手法の有用性を示している．

また，Google Scholar を利用した検索とも精度を比較する．Google Scholar も PubMed と同様にランキング形式で出力され，AND 検索を行うことができない．そこで，まず入力文献を検索した上で，その入力文献の関連記事を得て，その結果を PubMed の時と同様にスコア付けを行い，評価した．この結果，MAP は 0.66175 を得た．この評価は本研究によるものより高い値を示している．これについて，まず，Google Scholar によって得られる検索結果の数が本手法によるものと比べて大幅に少なく，このことが Google Scholar にとって非常に有利に働いている可能性がある．そこで，出力される数を Google Scholar と本手法のいずれか少ない方に合わせて評価した結果，MAP が本手法は 0.6386，Google Scholar は 0.6618 となり，依然本手法の方が若干，低い．しかしながら，Google Scholar は引用情報を元にランキングを判断していると考えられ，正解データを引用情報から得ている今回の方法では正当な評価になっていない可能性がある．どのような評価が妥当かについては今後の課題である．

具体的な検索結果について考察すると，例えば，入力セットが PDB ID:1c4z と PDB ID:1ayz のとき，本手法ではランキングの上位に PDB ID:1fbv を得ている．PDB ID:1c4z の title は “Structure of an E6AP-UbcH7 complex: insights into **ubiquitination** by the E2-E3 enzyme cascade.” であり，PDB ID:1ayz の title は “Crystal structure of the Saccharomyces cerevisiae **ubiquitin**-conjugating enzyme Rad6 at 2.6 $\AA$ resolution.” である．そして，PDB ID:1fbv の title は “Structure of a c-Cbl-UbcH7 complex: RING domain function in **ubiquitin**-protein ligases.” であり，“ubiquitin(ユビキチン)” に関する論文を得ている．この PDB ID:1fbv は Google Scholar では得られず，これは Google Scholar では得られなかった論文が，本手法では得られたことを示している．

4.3.3 フィルタリングの有用性

ここでは，前処理としてのフィルタリングが精度の観点，実行時間の観点から有用なものであるか検証する．表 3 の 21 種類の入力セットについて，フィルタリング前 (すなわち検索対象を PDB に登録された全てのデータとしたもの) と，フィルタリング後の結果の比較を図 4 に示す．図 4(a) は recall at top k のマクロ平均であり，図 4(b) は precision at top k のマクロ平均である．全ての入力セットに対する結果のうち，フィルタリングによって得られた文献数が最少となるものが 92 件であったため，ここでは 92 位までの平均値を算出している．

このグラフから，上位の結果に関しては Recall も Precision もフィルタリングを行った方が良い結果が得られている．これはフィルタリングにより関連の強い蛋白質を検索対象文献として残すことに成功し，下位にランキングされた文献が検索対象から外れていることを示している．検索システムにおいて上位にランキングされた結果が参照されやすいことを考えると，フィルタリングの精度は非常に良いものと考えることができる．

また，計算時間の観点について，メモリ 6GB，CPU は quad-core でクロック周波数 2.67GHz の Core i7 920 の計算機を一台用いて検索したところ，一回の検索に要する時間

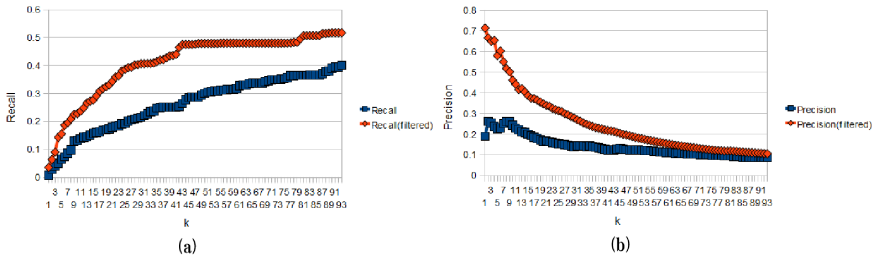

図 4 フィルタリング前後の Presicion at top k と Recall at top k
Fig. 4 Precision at top k and Recall at top k difference between no-filter and filtered

がフィルタリング前では平均 27.3 分，フィルタリング後では平均 8.2 秒であった．これより，フィルタリングを行うことによって計算時間を減らし，実用的な計算時間で検索することが可能となった．

5. ま と め

本研究では，蛋白質構造解析文献と複数の GO Term を関連付け，Gene Ontology の非循環有向グラフ上で概念間の経路から概念間の類似度を求め，これに基づく関連文献検索手法を提案した．検索システムを構築し評価した結果，(1) 概念的に蛋白質構造解析文献を検索することはテキスト検索に比べ有用な結果であり，(2) 付加文献を追加して Gene Ontology の非循環有向グラフに重み付けを行うことで，ユーザの意図を反映できることを示した．

一方，文献に掲載されている蛋白質に付与されている GO Term の数は蛋白質によって一定でないので，蛋白質によって類似度に対する GO Term 1 つ当たりの重要度が必ずしも同等ではない．このため GO Term の付与数が少ない蛋白質はその類似度が変化したときに大きくランキングが変化し，不安定になる可能性があり，それへの対処が望まれる．また，評価の観点からは，より多くの入力データに対する検索実験や参照情報を利用しないより適切な正解データセットの構築などが残された課題である．

参 考 文 献

1) Haiyuan Yu, Ronald Jansen, Gustavo Stolovitzky and Mark Gerstein:“Total ancestry measure: quantifying the similarity in tree-like classification, with genomic applications”, Bioinformatics, Vol. 23, no. 16, pp. 2163−2173, 2007
2) Shanfeng Zhu,Jia Zeng and Hiroshi Mamitsuka:“Enhancing MEDLINE document clustering by incorporating MeSH semantic similarity”, Bioinformatics, Vol 25, no. 15, pp. 1944−1951, 2009
3) Claudia Leacock and Martuin Chodorow:“WordNet: an electronic lexical database”, The MIT Press, pp. 265−283, 1998
4) National Center for Biotechnology Information (US):“PubMed Help”, Bethesda, pp. 26-27, 2005